ALPHA-HELICAL REPEAT PROTEIN FOLDING AND TURNOVER: A THERMODYNAMIC ANALYSIS OF NATURAL AND UNNATURAL REPEAT ARCHITECTURES

by

Kevin A. Sforza

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

August, 2017

Abstract

Studies using repeat proteins have been invaluable to our understanding of protein folding thermodynamics. Here we present work investigating the origins behind protein folding stabilities and protein turnover for designed alpha-helical repeat proteins. We have also developed a novel protein purification method exploiting polyaspartate and polyglutamate affinity tags binding on hydroxyapatite.

Consensus ankyrin repeat proteins are designed repeats with a naturally occurring structural conformation and remarkably high stability. We observe a direct relationship between levels of sequence conservation with changes in two-state folding stability revealed by circular dichroism (CD) spectroscopy with single-substituted repeats. In addition, one-dimensional Ising analysis revealed that consensus amino acids preferentially stabilize the repeat interfaces more than the intrinsic repeats. Consistent with previous NMR findings, we observed high interfacial stabilities conferred from residues involved in a buried polar network. Together, these findings support the importance of conserved buried interfacial polar residues in natural repeat protein conformations.

In contrast to natural repeat structures, applying Ising analysis to CD data for repeat proteins with unnatural conformations revealed stabilizing intrinsic repeat stabilities. The fast folding kinetics observed for these repeat proteins support a barrierless folding pathway. These

data uncover short-range interactions as the origins of extreme stability for these unnatural repeat architectures.

We investigated the relationship stability has with protein turnover using an *in vitro* assay to degrade consensus ankyrin repeats with *E. coli* Lon protease. While we observed an expected increase in rates of degradation with a decrease in global stability, however the distance-dependence of stability on proteolysis remains an open question.

We have developed a novel purification method exploiting polyaspartate and polyglutamate binding to hydroxyapatite. We observed binding independent of terminal placement, an increased affinity with longer tag length, and positive detection of polyglutamate tags by Western blot. Because the fusion-tag decreased rates of *E. coli* adenylate kinase catalysis, we recommend tag removal from active enzymes.

Overall, these results provide insight into the relationship between protein sequence and stability for a variety of protein architectures. It is our hope that the findings and molecular tools developed here will prove useful for future protein engineering studies.

Thesis advisor: Dr. Douglas E. Barrick

Second reader: Dr. Vincent J. Hilser

Thesis committee: Dr. Juliette T.J. Lecomte

Dr. Gregory D. Bowman

To my grandfather,

Richard Sforza,

who taught me to believe in magic.

Acknowledgements

The following body of work details the “whats” and “hows” of protein folding thermodynamics as they relate to my personal quest to chip away at life’s “whys”. This page is dedicated to the all-important “whos”, whose support, patience, and compassion have allowed me to unapologetically pursue my dreams.

To Dr. Douglas E. Barrick, my mentor and real-life wizard – You have a curiosity that is unmatched by fiction. Your ideas and critiques transform the ways in which we know the world. If a day ever comes where you stop dreaming, please leave me out of the loop. You talk about science and statistics like chefs do about butter. Some of my favorite conversations with you have been when we step back from our data and weave them into to the bigger picture. Thank you for taking me under your wing. I am forever grateful for the opportunities you have given me to grow in your lab.

To Drs. Juliette T.J. Lecomte, Vincent J. Hilser, and Gregory D. Bowman, my thesis committee – Thank you for pushing me to question everything, just as each of you has questioned everything I have shared with you. Your guidance and objective criticism helps keep me honest, and I am very grateful for the ideas and wisdom you have imparted on me over the years.

To The Johns Hopkins University Biophysics and Biology faculty and staff – You have moved mountains for me over the years, and yet you still continue to say "yes" when I ask for more. Thank you for welcoming me to Baltimore and providing me with a home away from home.

To my labmates and science siblings – Without you, the ebbs and flows of graduate school could have beaten past as gray doldrums. Our lab and meetings provide a haven for constructive criticism where we peel back the curtains to gawk at the "emperor's new clothes." To Dr. Katie Tripp, Christine Hatem, Dr. Eva Cunha, Dr. Thuy Dao, Dr. Jake Marold, Dr. Kate Sherry, Katie Geiger-Schuller, Sean Klein, Matt Sternke, and Mark Peterson – words cannot express how grateful I am to have learned with you and from you over the years. I am privileged to call you my colleagues, but most importantly my friends. You breathe life through your pipets and I am excited to watch you change the world.

To my family – "SF" is the trickiest consonant combination I have ever had the pleasure to watch other people slip over. My life has been a series of fortunate events because of your endless love, and I can always count on you for sound guidance when I fall down. To Dad, Mom, Chris, Brian, Alicia, grandparents, aunts, uncles, and cousins – you are each caring, funny, forgiving, and kind. Thank you for always accepting me for who I am.

To each of my friends – Stochastic variations brought you into my life, but things unseen keep us bound together. We return to very

different lives after we get together, but the impact you have made on me leaves a lasting impression. Thank you for making me laugh and cry, but most importantly thank you for never letting me down.

To my fiancé Anthony – We speak our own language and I know everyone else thinks we are crazy. Words fail to express how happy I am to be sharing our lives together. One day we are going to look back fondly at the great times we had with our puppy in that tiny one-bedroom on Saint Paul St. Thank you for your unyielding support and optimism. I cannot wait to see the great things we accomplish together.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Natural selection and the limits of protein folding

American automobile titan Henry Ford is touted for his economical and industrial efficiency. In his 1976 essay, Humphrey highlights this efficiency in describing Ford's reverse engineering of an automobile engine's kingpins, stating:

> "Henry Ford, it is said, commissioned a survey of the car scrap yards of America to find out if there were parts of the Model T Ford which never failed. His inspectors came back with reports of almost every kind of breakdown: axles, brakes, pistons -- all were liable to go wrong. But they drew attention to one notable exception, the kingpins of the

> scrapped cars invariably had years of life left in them. With ruthless logic Ford concluded that the kingpins on the Model T were too good for their job and ordered that in future they should be made to an inferior specification" (Humphrey: p303 of Bateson et al., 1976)

In his 2011 article, David Mikkelson discusses how this tale is most likely anecdotal and has never been substantiated by any other sources close to Henry Ford during his lifetime (Mikkelson, 2011). Mikkelson goes on to evaluate that some modern scientists have used it as a metaphor for natural selection (Barrow, 1996; Dawkins, 1996; Diamond, 1997) proposing that nature does not require mechanisms to be more efficient than necessary. Although Darwinian evolution selects for survival advantages over time, the mechanisms for short-term natural selection do not have the ability to favor the efficient over the just barely-competent. Here we extend this formalism into the realm of protein folding.

For most proteins, stretches of polypeptides fold into beautiful three-dimensional structures, known as the "native" state. A polypeptide of length n, has n-1 peptide linkages, and a total of 2(n-1) phi and psi angle values. Depending on the number of possible values that each of these angles can adopt, the number of folded conformations rises at an

explosive rate. In spite of this, proteins spontaneously fold into their correct conformations with high degrees of precision and speed, lending credence to Levinthal's paradox. It is astonishing that proteins regularly fold correctly, and that life continues to thrive. In most cases, a protein's folded structure is dictated solely by the information encoded by its amino acid sequence. This was originally outlined by Christian B. Anfinsen's Nobel Prize winning research on ribonuclease A (Anfinsen, 1973). In this work, Anfinsen postulates that the protein native state has the lowest (and thus most favorable) Gibbs free energy out of all the available backbone conformations.

By contemplating the thermodynamic limits of protein folding, we can better evaluate the range of possible free energies. For folded proteins under biological conditions, the upper free energy limit (lower stability limit) is anything less than the value of zero (i.e., negative), where the folding equilibrium favors the folded state. However, the lower limits of folding free energy (i.e., the upper limits of stability) are less clear. Besides favoring the folded state, what does it mean for proteins to have a highly negative folding free energy? Why don't we observe the full gambit of stabilities in nature, from folds that just barely adopt stable conformations, to super-stable folded structures that resist access to the unfolded state?

In Chapter 2 of this work we examine the sequence determinants of a consensus designed repeat protein. While such consensus designed proteins are derived from naturally occurring fold families, and faithfully maintain their native folds, they have thermodynamic stabilities that surpass their naturally occurring counterparts. We find that the conservation levels of different residues are directly related to the thermodynamic stability they impart. Next in Chapter 3 we examine the folding stabilities of a series of designed repeat proteins whose geometries are to-date unobserved in nature. We find that these designed repeat proteins differ from those that exist in nature, displaying favorable intrinsic free energies of folding. Through the use of an *in vitro* cellular protease assay in Chapter 4, we begin to dissect the extent to which protein stability dictates rates of degradation. We find that rates of proteolysis increase as a direct consequence of destabilization. Finally, in Chapter 5 we have developed novel affinity tag to be used in the purification of recombinant proteins. It is our hope that the use of this purification protocol will alleviate some of the problems laboratory scientists in generating high quantities of purified proteins.

Through this work, we find that nature has selected for sequences that are just favorable enough to result in the correct folded structures. Unlike the consensus and Rosetta designed proteins discussed here, nature does not have the ability (or the need) to over-engineer proteins to be more stable than necessary. In fact, if the aforementioned Henry Ford

anecdote had been applied to protein sequence design, we would most likely have the same sequences and conformations we observe today. Just as the kingpins do nothing to enhance Ford's Model T on their own, exceptional stability imparts little (if any) fitness advantage. Neither the kingpins nor the proteins should be expected to be any better than average. In fact, if protein folding stabilities in nature were too high, cellular processes requiring unfolding would be severely hindered.

1.2 Linear repeat proteins and their utility in studying protein folding thermodynamics

Protein folding studies often exploit the controlled titration of denaturants to unfold proteins and perturb the folding equilibrium. In ensemble-averaged equilibrium experiments, changes are made to protein primary structure or solution conditions to investigate the folding behavior of proteins at the global level.

To resolve differences in higher order contacts between elements of globular protein structure, domains and subdomains of the intact protein are removed. This means of investigation imposes many different

challenges to folding studies. Interacting amino acids in a protein's three-dimensional folded space are not necessarily close in primary structure. For most globular proteins, residues interact with other residues in both a sequence-close and sequence-distant manner. Deletion of regions of primary structure close to one terminus of the protein primary structure can result in local and global disruptions to the folded state of the remaining protein. Because of this, many deletion mutants are prone to misfolding, unfolding, and aggregation, hindering equilibrium folding experiments (Chow et al., 2003).

Repeat proteins are successfully used as a simplified experimental system in order to avoid the experimental limitations imposed by globular proteins. Linear repeat proteins are composed of repeating modular units of secondary structure that extend in a single direction away from the protein's N-terminus. Owing to their linear organization, both the local and global interactions observed for each residue are relegated to residues within the same modular repeat and between adjacent repeats. This linear architecture allows for the successful addition and deletion of structural units.

There are many different types of linear repeat proteins that are composed of single or mixed α-helices (Figure 1.1A), β-sheets (Figure 1.1B), and coiled structural elements (Figure 1.1C). Each type of repeat provides different surface structures that modulate a variety of protein-

mediated interactions. For a review of the major repeat protein families and their functions, see Andrade et al., 2001; Kajava, 2001. In this study, we utilize the α-helical repeat proteins, consensus ankyrin repeat proteins and designed helical repeat proteins (DHRs), to probe questions relating to protein sequence, structure, and thermodynamic stability.

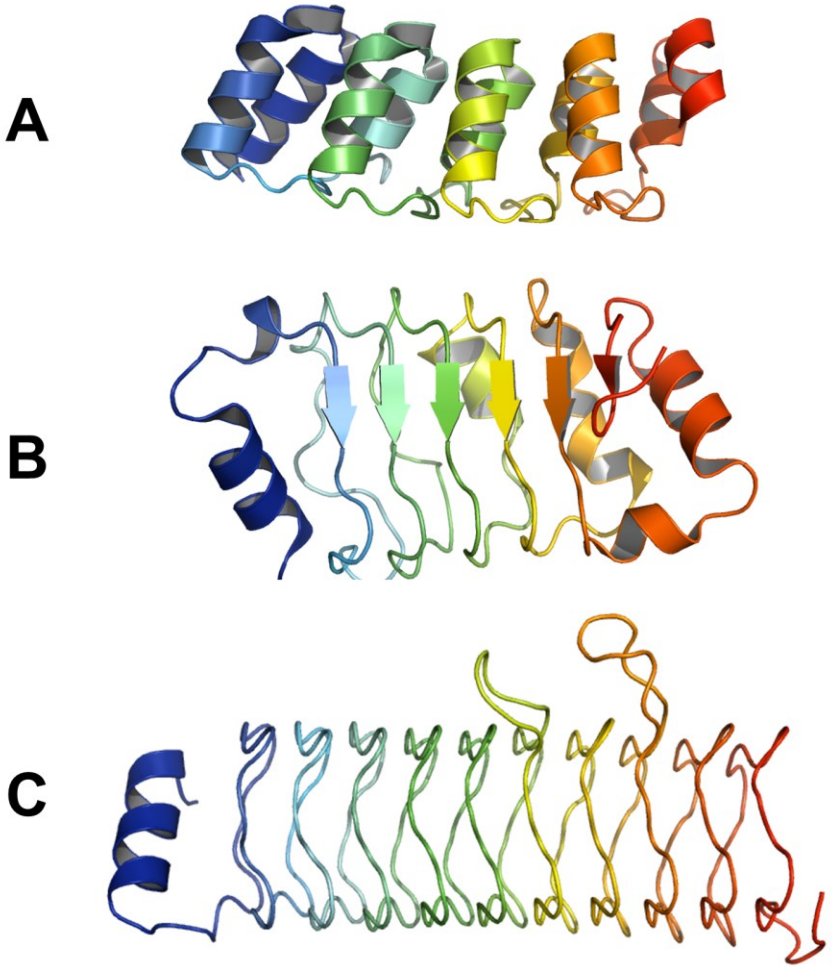


Figure 1.1 The conformational diversity of linear repeat proteins. (A) Crystal structure of designed consensus ankyrin repeat protein E3_19 (PDB: 2BKG). (B) Crystal structure of the tumor suppressor leucine rich repeat protein PP32 (PDB: 2JE1). (C) Crystal structure of the HetL pentapeptide repeat protein (PDB: 3DU1). Tandem repeats form linear protein arrays of varying secondary structural elements, shown in rainbow ribbon from N-terminus (left, blue) to C-terminus (right, red).

1.3 Consensus protein design

Although naturally occurring repeat proteins maintain structural similarity from one repeat to the next, sequences within a particular repeat type vary greatly. To further simplify the study of repeat proteins, natural variation across repeats can be averaged through the use of consensus design.

Consensus proteins are designed through the analysis of multiple sequence alignments where the most frequent amino acid is used at each position of the protein sequence. While a consensus sequence of related proteins can be easily identified, the sequences found in the multiple sequence alignment vary significantly both from one another, and from the consensus sequence. Furthermore, because consensus protein design utilizes only the most statistically probable amino acid at each position, covariance information between interacting residues can be lost. Understanding the statistical interactions between amino acid positions is useful in protein folding thermodynamics (Lockless and Ranganathan, 1999). However, these interactions are captured only if the frequencies of the covariant amino acids are the most probable in the alignment.

Consensus designed repeat proteins have previously been shown to have higher thermodynamic stability compared to naturally occurring

proteins (Kajander et al., 2006; Tripp and Barrick, 2007; Wetzel et al., 2008). However, the nature of this stark stability increase is unclear, especially in light of the fact that the consensus sequence is designed from its naturally occurring counterparts. In Chapter 2 we begin to address the open question regarding the nature of the relationship between protein sequence and stability as it relates to consensus protein design.

1.4 Determining the number of repeat constructs necessary for Ising analysis

Ising methodologies for analyzing linear repeat proteins have been previously described in great detail (Aksel et al., 2011; Geiger-Schuller and Barrick, 2016; Marold et al., 2015) and are applied to different linear repeat proteins in Chapter 2 and Chapter 3 of this work. All of these works utilize a matrix formalism to describe the statistical weights for each repeat. These statistical weights are combined to form the full partition function for each repeat protein used in the Ising determination of contributions to folding free energy. However, this approach describes how the fitted nonlinear model is constructed, but offers little insight into to how scientists determine the number of constructs necessary to perform Ising analysis. One way to think about this

problem is to represent global free energies of an n-repeat protein as a linear sum of intrinsic (i) and interfacial (i, i+1) contributions:

$$\Delta G^\circ = \sum_{i=1}^{n} \Delta G^\circ_i + \sum_{i=1}^{m} \Delta G^\circ_{i,i+1} \quad (1.1)$$

$$[\Delta G^\circ] = [n \quad m] \bullet \begin{bmatrix} \Delta G^\circ_i \\ \Delta G^\circ_{i,i+1} \end{bmatrix} \quad (1.2)$$

Selecting the right constructs for analysis is analogous finding conditions to reliably solve a system of linear equations like Equation 1.1. Shown above (Equation 1.2) is the simplest matrix form of this system of equations. We can describe the free energy contributions for a three repeat protein of type R (i.e., RRR) using three intrinsic free energies, and two interfacial free energies. This is shown below in Equation 1.3, and by the dot product form in Equation 1.4.

$$[\Delta G^\circ(RRRR)] = [3(\Delta G^\circ_i) \quad 2(\Delta G^\circ_{i,i+1})] \quad (1.3)$$

$$[RRRR] \bullet [\Delta G^\circ] = [3 \quad 2] \bullet \begin{bmatrix} \Delta G^\circ_i \\ \Delta G^\circ_{i,i+1} \end{bmatrix} \quad (1.4)$$

Repeat proteins of the same repeat-type are sometimes prone to aggregation, as is observed with consensus ankyrin repeats (Aksel et al., 2011). To combat this, solubilizing substitutions are made to N- and C-terminal capping repeats. For proteins with N- and C-terminal solubilizing caps, we need to define additional parameters related to the sequence differences in the caps. This is done if we are assuming different free energy parameters for each repeat. Shown below (Equation 1.5) is a formula for describing the free energy contributions for a four repeat protein NR_2C with different intrinsic and interfacial terms. Notice that for this representation, there are six unknown parameters: the intrinsic terms for the N-, R-, and C-repeats, and interfacial terms for the N-R, R-R, and R-C interfaces.

$$[NRRC] \bullet [\Delta G^\circ] = [1 \quad 2 \quad 1 \quad 1 \quad 1 \quad 1] \bullet \begin{bmatrix} \Delta G^\circ_N \\ \Delta G^\circ_R \\ \Delta G^\circ_C \\ \Delta G^\circ_{N-R} \\ \Delta G^\circ_{R-R} \\ \Delta G^\circ_{R-C} \end{bmatrix} \tag{1.5}$$

Because we define six unique terms, at bare minimum we would require six unique sets of data of different construct type and length to resolve those intrinsic and interfacial terms. Below (Equation 1.6) we

demonstrate this with one possible matrix solution for repeat proteins of type R with N- and C-terminal repeats.

$$\begin{bmatrix} NR \\ RC \\ NRR \\ RRC \\ NRC \\ NRRC \end{bmatrix} \bullet [\Delta G^\circ] = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 2 & 0 & 1 & 1 & 0 \\ 0 & 2 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 \end{bmatrix} \bullet \begin{bmatrix} \Delta G^\circ_N \\ \Delta G^\circ_R \\ \Delta G^\circ_C \\ \Delta G^\circ_{N-R} \\ \Delta G^\circ_{R-R} \\ \Delta G^\circ_{R-C} \end{bmatrix} \tag{1.6}$$

When determining the number of constructs necessary to resolve each free energy parameter, the rank of the coefficient matrix cannot be less than the number of fit parameters. The row rank of a matrix is the maximum number of linearly independent rows in the matrix. If one row or column is a multiple of another, then they are not independent and the determinant is zero. Therefore, additional repeat protein constructs of greater length or varied capping structure are required to obtain full rank. We have applied these principles in Chapter 3 to determine the Ising fit parameters for designed helical repeat proteins.

These principles can be extended to different types of repeat proteins. In Chapter 2, we introduce a series of point substitutions to consensus ankyrin R-repeats. For these substituted repeats, which we denote as R*-repeats, we determine intrinsic R* free energies, as well as N-R*, R-R*, R*-R*, R*-R, and R*-C interfacial free energies. Because

previous work already determined the Ising parameters for R-, N-, and C-repeats (Aksel et al., 2011), we treat those values as constants and fit only parameters that are perturbed by point substitutions.

1.5 Protein stability modulates mechanisms of protein degradation.

The ability of cells to degrade proteins not only replenishes pools of free amino acids for use in the synthesis of new polypeptides, but over the course of evolutionary time has served as a fundamental bottleneck exploited by natural selection. Bacterial proteases are most likely the earliest forms of these machines in evolutionary time (Goldberg, 1972), whereas eukaryotic ubiquitin-dependent proteases evolved more recently (Glickman and Ciechanover, 2002).

Promiscuous and substrate-specific proteases play a variety of roles in the lifetimes of proteins and cells. Components of the cellular proteasome and intracellular proteases regularly degrade functional proteins in an effort to regulate and turn on and off protein-mediated pathways. Cell cycle and cell division checkpoint cyclins are degraded in order for the cell to proceed through rounds of replication (Glotzer et al., 1991), caspases involved in apoptosis are activated by the proteolytic

cleavage of procaspases (Thornberry and Lazebnik, 1998), and cholesterol biosynthesis and uptake is regulated through the cleavage of membrane-bound SREBP (Horton et al., 2002).

Proteases also act as cellular housekeepers to degrade and prevent the buildup of abnormal and misfolded proteins. Proteases owe their ability to regulate protein turnover partly to the fact that protein structures fluctuate throughout the ensemble of native folded conformations (Bai et al., 1995). Regions of local instability are shown to be more prone to proteolysis than those of higher stability (Park and Marqusee, 2005). In times of cellular stress like heat shock, proteins access conformations distinct from their native ensemble and are rapidly degraded to avoid aggregation (Parsell and Lindquist, 1993). Many human neurodegenerative diseases display a protein aggregation phenotype. Alzheimer's, Parkinson's, Huntington's, cystic fibrosis, and prion diseases are just a few of such proteopathies consistent with decreased proteasome activity.

Because of their indispensable regulatory functions, proteases most likely play a major role in natural selection. If the degradation of a protein promotes survival of an organism, that organism will have a selective advantage over other members of the same cohort lacking that mechanism for degradation. Contrary to this, if a misfolded protein promoting aggregation evades degradation machinery, then the host cell

or organism might have a selective disadvantage over other members of the same cohort.

In Chapter 4, we begin to examine the role local stability changes impart on proteolytic degradation by a processive protease. We use consensus ankyrin repeats to explore the extent to which local perturbations to stability affect global rates of degradation, and to explore the distances over which they act.

1.6 References

Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993 19, 349–360.

Andrade, M.A., Perez-Iratxeta, C., and Ponting, C.P. (2001). Protein repeats: structures, functions, and evolution. J. Struct. Biol. 134, 117–131.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science 181, 223–230.

Bai, Y., Sosnick, T.R., Mayne, L., and Englander, S.W. (1995). Protein folding intermediates: native-state hydrogen exchange. Science 269, 192–197.

Barrow, J.D. (1996). The Artful Universe: The Cosmic Source of Human Creativity (Boston: Back Bay Books).

Bateson, P.P.G., Hinde, R.A., and Humphrey, N.K. (1976). Growing Points Ethology (Cambridge: Cambridge University Press).

Chow, C.C., Chow, C., Raghunathan, V., Huppert, T.J., Kimball, E.B., and Cavagnero, S. (2003). Chain Length Dependence of Apomyoglobin Folding: Structural Evolution from Misfolded Sheets to Native Helices. Biochemistry (Mosc.) 42, 7090–7099.

Dawkins, R. (1996). River Out of Eden: A Darwinian View of Life (New York, NY: Basic Books).

Diamond, J. (1997). Why is sex fun? : the evolution of human sexuality (New York NY: HarperCollins).

Geiger-Schuller, K., and Barrick, D. (2016). Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. Biophys. J. 111, 2395–2403.

Glickman, M.H., and Ciechanover, A. (2002). The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol. Rev. 82, 373–428.

Glotzer, M., Murray, A.W., and Kirschner, M.W. (1991). Cyclin is degraded by the ubiquitin pathway. Nature 349, 132–138.

Goldberg, A.L. (1972). Degradation of Abnormal Proteins in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 69, 422–426.

Horton, J.D., Goldstein, J.L., and Brown, M.S. (2002). SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. J. Clin. Invest. 109, 1125–1131.

Kajander, T., Cortajarena, A.L., and Regan, L. (2006). Consensus Design as a Tool for Engineering Repeat Proteins. In Protein Design, (Humana Press), pp. 151–170.

Kajava, A.V. (2001). Review: proteins with repeated sequence--structural prediction and modeling. J. Struct. Biol. 134, 132–144.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286, 295–299.

Marold, J.D., Kavran, J.M., Bowman, G.D., and Barrick, D. (2015). A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Struct. Lond. Engl. 1993 23, 2055–2065.

Mikkelson, D. (2011). Henry Ford Junkyard Parts. http://www.snopes.com/business/genius/fordpart.asp

Park, C., and Marqusee, S. (2005). Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. Nat. Methods 2, 207–212.

Parsell, D.A., and Lindquist, S. (1993). The function of heat-shock proteins in stress tolerance: degradation and reactivation of damaged proteins. Annu. Rev. Genet. 27, 437–496.

Thornberry, N.A., and Lazebnik, Y. (1998). Caspases: enemies within. Science 281, 1312–1316.

Tripp, K.W., and Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. J. Mol. Biol. 365, 1187–1200.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. 376, 241–257.

Chapter 2

Stability of consensus ankyrin repeat protein is directly related to information content

2.1 Introduction

Repetitive protein sequences are ubiquitous across the three main domains of life. Repeat sequences are thought to arise by gene duplication events over the course of evolutionary time (Pascual et al., 1997; Schüler and Bornberg-Bauer, 2016). Unlike their globular counterparts whose complex structure allows for long-range interactions, repeat protein structure is dominated primarily by interactions that are

short-range in sequence. These contacts primarily occur either within the repeat itself or between adjacent repeats.

Ankyrin repeat proteins are one of the most common types of repeat protein domains whose sequences pervade most forms of life. Ankyrin repeat are 33 residues, and form a structural helix-loop-helix motif. The repetitive alpha-helical structure of ankyrin repeats permits structural characterization through far-UV circular dichroism spectroscopy. By monitoring chemical and thermal unfolding transitions, ankyrin repeat proteins have proven to be a robust system to study protein folding thermodynamics (Aksel et al., 2011; Tripp and Barrick, 2007, 2008; Zweifel et al., 2003).

Repeat-to-repeat sequence variation in natural proteins complicates structural and thermodynamic studies. To simplify this, a homologous consensus sequence was successfully designed using the most frequent amino acid at each position of the ankyrin multiple sequence alignment (Binz et al., 2003; Mosavi et al., 2004; Tripp and Barrick, 2007). N- and C-terminal capping repeats were designed by the substitution of four polar residues into putative solvent accessible hydrophobic consensus residues (Aksel et al., 2011). The consensus ankyrin repeat sequence is referred to as repeat “R” with designed solubilizing “N” and “C” capping repeats (Figure 2.3B).

A hallmark of most consensus designed proteins is a marked increase in thermodynamic stability (Kajander et al., 2006; Komor et al., 2012; Tripp and Barrick, 2007; Wetzel et al., 2008). Previous work using nearest-neighbor Ising analysis to determine local stability and long-range cooperativity has uncovered the origins of the increased stability and cooperativity observed in consensus ankyrin repeat proteins. The folding of individual repeats provides a thermodynamically unfavorable free energy (5.24 ± 0.17 kcal*mol^{-1}) while the pairing of adjacent folded adjacent repeats provides a highly favorable (-12.54 ± 0.27 kcal*mol^{-1}) interfacial free energy (Aksel et al., 2011). Thus, the addition of each additional repeat provides a free energy decrease of -7.3 ± 0.32 kcal*mol^{-1}.

By comparing the global stability of naturally occurring ankyrin repeat proteins of increasing number of repeats with consensus ankyrin repeat proteins of corresponding length (Figure 2.1), it is immediately apparent that the trend in added stability conferred by adding a single consensus repeat is not shared by naturally occurring repeat proteins. This stark difference in thermodynamic behavior between naturally occurring protein sequences and consensus protein sequences, which are designed from their natural homologues, brings about the basic question – why are consensus proteins strikingly more stable than the natural sequences that gave rise to them?

Here we construct a series of consensus ankyrin repeat proteins of different lengths with a series of natural amino acid substitutions away from consensus, to test how closely conservation tracks with thermodynamic stability. Guanidine-HCl induced unfolding transitions of these consensus ankyrin repeat variants monitored by far-UV circular dichroism are resolved into their contributions to protein folding stability. We compare the stability changes conferred by these substitutions to the log probabilities in multiple sequence alignment of naturally occurring ankyrin repeats and observe a direct relationship between the two.

2.2 Results

2.2.1 Sequence entropy of ankyrin repeat proteins.

To date, the full available sequence alignment for the ankyrin repeat protein domain (Pfam entry: Ank (PF00023)) contains 8237 available sequences across 1099 species. Though some ankyrin repeat proteins appear in bacteria and archaea, they are far more common in eukaryotes. The Pfam database full sequence alignment contains all detectable sequences related to the ankyrin family of sequences. This alignment, while extensive, contains exact duplicate entries and is not well-maintained. For each protein family, Pfam maintains seed alignments that contain a small set of representative sequences that are carefully curated. For this reason, we have used the Pfam seed alignment for our hidden Markov model (HMM) logo (Wheeler et al., 2014) and sequence entropy calculations (Figure 2.3 A,C).

The HMM logo for the ankyrin multiple sequence alignment shows that the most probable residue at each position, the top letter, contributes the largest amount of information content to that position in the alignment. In Figure 2A at each position of the multiple sequence

alignment, the height of the stack of letters, or the total information content (IC), is equal to the Kullback-Leibler distance, as defined by Equation 2.1, where p_i is the frequency of an amino acid in the alignment, and q_i is the background frequency of that amino acid in the distribution (Wheeler et al., 2014). Within each stack of letters, the height of the individual letters are proportional to their individual frequencies. There are 11 differences between the consensus sequence predicted from the top letters of the HMM logo (positions 3, 11, 12, 14, 15, 17, 23, 24, 28, 31, 33) and the current consensus R-repeat sequence (Figure 2.3 B).

$$\mathrm{IC} = -\sum_i^n p_i log_2(\frac{p_i}{q_i}) \quad (2.1)$$

To evaluate the levels of sequence variation across the multiple sequence alignment, we determined the relative levels of Shannon entropy, H(x), at each position of the multiple sequence alignment using Equation 2.2. At a given position in a multiple sequence alignment, the Shannon entropy is determined by taking the negative of the product of the probability (p_i) of an amino acid, i, and the log base 2 of that probability, summed over all 20 amino acids, n, at that position (Shannon, 1948). Shannon entropy (Equation 2.2) differs from

information content (Equation 2.1) only in that it is not scaled by the background distribution of amino acid frequencies, q_i.

$$H(x) = -\sum_i^n p_i log_2(p_i) \quad (2.2)$$

Shannon entropy is a measure of uncertainty of an outcome. In this case, the Shannon entropy at any position is a measure of the uncertainty of the identity of an amino acid at any position of a multiple sequence alignment. Entropy is the greatest when amino acid identity is the most uncertain at a position (when equal frequencies of all 20 amino acids are present p_i=0.05 for all residues and H(x) approaches 4.3), and it reaches a minimum when amino acid identity is most certain at a position (when only one amino acid is present in an alignment, H(x) approaches 0). Figure 2.3C shows a plot of sequence entropy for each position of the ankyrin multiple sequence alignment.

The amino acid residues that differ between the consensus R-repeat and the HMM logo all have relatively high sequence entropy levels, reflecting a fairly high level of uncertainty of amino acid identity at that position. Qualitatively, the HMM logo also reflects the level of uncertainty revealed by the Shannon entropy. At positions of the HMM logo where the letter height between the top-most and the lower letters is

indistinguishable, the Shannon entropy is maximal and uncertainty is greatest.

When the position on the HMM logo appears to be dominated by a single letter whose height towers over the landscape (positions 2, 4, 5, 6, 7, 9, 21, and 22), Shannon entropy is lowest, reflecting relatively low uncertainty in the alignment. It should be noted that there are some positions (positions 25, 26) that have relatively high sequence identity according to the HMM logo, while still containing high levels of Shannon entropy and uncertainty. This is due to the highly variable and equally low proportioned probabilities of residues beneath the highest, most probable letter in the HMM logo at that position.

Based on the HMM logo and Shannon entropy, we chose four positions to determine the effects these substitutions away from the consensus sequence have on protein stability. The positions 4, 6, and 21 were chosen because of their relatively high levels of sequence conservation, as marked by the height of their letters in the HMM logo and low levels of Shannon entropy (Figure 2.3 A,C). Position 28 was chosen because of the highly variable residue identity at this position, and its identity in the consensus R-sequence no longer matched the highest level in the HMM logo. The Shannon entropy at position 28 is relatively high, indicative of high identity uncertainty and poor sequence conservation.

While sequence entropy is a useful global metric for examining sequence conservation across all amino acids at a position in an alignment, we would like to examine the individual contribution each amino acid has to the sum total of the alignment. Assuming the amino acid distribution in the multiple sequence alignment follows the Boltzmann distribution (Equation 2.3) where p_i is the probability of an amino acid, i, in the multiple sequence alignment, G_i is the free energy of residue i at that position in the protein sequence, T is temperature, and k_B is the Boltzmann constant. Equation 2.3 can be rearranged to give Equation 2.4 for free energy at residue i in a multiple sequence alignment.

$$p_i = e^{-G_i/k_BT} \tag{2.3}$$

$$G_i = -k_BTln(p_i) \tag{2.4}$$

For a protein sequence from a multiple sequence alignment, each residue contributes to the probability distribution of each position of the alignment. When a point substitution is made, we define a term for the difference in contribution to the free energy of the amino acid distribution in the multiple sequence alignment, ΔG_{MSA} (Equation 2.5)

where p_f is the probability of the substituted residue, and p_i is the starting residue in the protein sequence. For consensus ankyrin, we define ΔG_{MSA} using the probability of the substituted residue relative to the consensus residue (Equation 2.6).

$$\Delta G_{MSA} = G_f - G_i = {}^{-}k_B T ln(p_f) - {}^{-}k_B T ln(p_i) \quad (2.5)$$

$$\Delta G_{MSA} = {}^{-}k_B T ln(p_{Substitution}) - {}^{-}k_B T ln(p_{Consensus}) \quad (2.6)$$

For each of the four substitutions ΔG_{MSA} values are relatively high (Table 2.1), reflecting their differences in sequence conservation. The substitution V28P, a position of high Shannon entropy and uncertainty has the lowest ΔG_{MSA}.

2.2.2 Guanidine-HCl induced unfolding transitions of consensus ankyrin repeat proteins.

To determine the effects that point substitutions away from the consensus R sequence have on overall fold stability, we monitored guanidine-HCl induced unfolding using circular dichroism spectroscopy. We first measured the folding stability of consensus ankyrin repeat proteins of 3- and 4-repeats, NRC and NRRC respectively (Figure 2.2). The two constructs unfold by single sigmoidal transitions, which are fit well with a two-state model. Comparing the data and fits for NRC and NRRC respectively, there is a large increase in protein stability with the increase in repeat number, consistent with previous findings (Aksel et al., 2011). The global free energy of folding, ΔG°, and m-value for NRC are -5.7 ± 0.19 kcal*mol^{-1} and 2.33 ± 0.05 kcal*mol^{-1}*M^{-1} respectively, and for NRRC the fit ΔG° and m-value are -10.84 ± 0.54 kcal*mol^{-1} and 2.69 ± 0.09 kcal*mol-1*M^{-1}. We use these values for stability comparisons with consensus ankyrin repeats containing substitutions away from consensus.

Using the three- and four-repeat ankyrin constructs, NRC and NRRC respectively, we made substitutions to the central R repeats; we represent substituted central repeats as R*. To determine the extent to which the consensus protein context affects changes in folding stability, we compared constructs where the substitution is either in the first or second central repeat of NRRC (NRR*C or NR*RC), or both central repeats, NR*R*C.

The unfolding transitions of these variants are also well-fitted using a two-state model (Figure 2.4). The parameters obtained from these two-state fits reveal the changes in global stabilities (Table 2.2). Comparing different series of point substitutions, it is apparent that each substitution affects global fold stability to varying degrees. When the substitution is clearly destabilizing, having two substituted repeats compounds the destabilizing effect, variant L6I for example (Figure 2.4). However, this stability change is not always linearly proportional to the number of protein repeats or the number of substituted repeats, variant T4V for example (Figure 2.4).

We observe a linear relationship when comparing the changes in folding free energy between consensus ankyrin repeats and the consensus ankyrin variants (pearson correlation coeff of X, p=Y), $\Delta\Delta G$, with differences in contribution to the multiple sequence alignment free energy, with a slope of X and an intercept of Y, ΔG_{MSA}, (Figure 2.5). This

observation directly relates the amount of information each residue has in the alignment with the stability changes observed in the consensus background. Substitutions in multiple repeats within the same protein compound the observed destabilizing effect.

2.2.3 Ising analysis of guanidine-HCl induced unfolding transitions.

Previous work has determined the intrinsic and coupling folding contributions to the folding stability of consensus ankyrin repeat proteins (Aksel et al., 2011). Fitted stability parameters obtained from this work determined that the intrinsic fold free energies of the N, R, and C repeats are all thermodynamically unfavorable at 6.1 ± 0.17 kcal*mol^{-1}, 5.2 ± 0.17 kcal*mol^{-1}, and 7.8 ± 0.19 kcal*mol^{-1} respectively. However, the unfavorable intrinsic stability is compensated by a large favorable coupling free energy of -12.5 ± 0.27 kcal*mol from the interfaces between folded repeats.

To determine the intrinsic and nearest-neighbor coupling free energies for the four variants studied here, we globally fit an Ising model to the to the guanidine-HCl induced unfolding transitions in Figure 2.6.

Due to the sequence, structure, and resulting stability differences presented by the N- and C-terminal capping repeats, additional R*RC and NRR* constructs were added to each set of melts.

In our fits, we set the intrinsic fold stabilities of N, R, and C to be constant values from previous work (Aksel et al., 2011), and assume the N-R and R-C interfaces to be equal. We then fit for the intrinsic fold stability of the substituted repeat, $\Delta G^{\circ}_{R^*}$, as well as the different interfaces repeat R* makes with the consensus ankyrin repeats, $\Delta G^{\circ}_{N\text{-}R^*}$, $\Delta G^{\circ}_{R\text{-}R^*}$, $\Delta G^{\circ}_{R^*\text{-}R}$, $\Delta G^{\circ}_{R^*\text{-}C}$. We also fit a unique interface between two substituted consensus repeats, $\Delta G^{\circ}_{R^*\text{-}R^*}$. We then fit a single m-value parameter to account for the effect of denaturant on intrinsic (but not interfacial) folding free energy.

The Ising model fits well to the guanidine-HCl unfolding transitions with 7 shared parameters (Table 2.3) fit to 6 curves from different protein constructs in triplicate (Figure 2.6). To calculate the uncertainty of the fit parameters at the 95% confidence interval, bootstrap analysis of the residuals was performed over 1000 iterations. All variants still exhibit the characteristic trend of the consensus mother protein of unfavorable intrinsic folding coupled to a much greater favorable interfacial stability. The variants studied affect the unfavorable intrinsic folding free energies to varying degrees (Figure 2.7A). However, we observe that all of the

studied substitutions away from consensus decrease the highly favorable interfacial free energies between two R* substituted repeats (Figure 2.7B).

The threonine residue at position 4 has been shown in crystal structures (PDB 2BKG, 1N0R) to play a role in a buried hydrogen-bond ladder through the core of the ankyrin repeat protein (Figure 2.8A), stabilizing the overall fold architecture. Through solution-NMR experiments, it has been confirmed that the threonine hydroxyl acts as a proton donor in a bifurcated hydrogen bond with a C-terminal histidine Nδ at position 7 within the same repeat (Preimesberger et al., 2015). The isosteric replacement of threonine with valine resulted in a decrease of approximately 0.39 ± 0.22 kcal*mol^{-1} in intrinsic stability, possibly due to the expected decrease in stability for a loss of a bifurcated hydrogen bond.

Furthermore, the Nε of the same histidine hydrogen bonds with the backbone carboxyl group of the residue at position 3 of the next C-terminal repeat. Overall, the T4V interfacial free energy terms show the trend that removing the threonine hydroxyl group to another repeat decreases the stabilizing interactions. The strongest destabilizing effect is observed for the R*-R* interface, followed by the R*-R and R-R* interfaces. In contrast, there is no significant change in stability of the R*-C interface. This is consistent with previous findings of decreased J-

coupling between the C-terminal H-bonding partners compared to those of internal repeats (Preimesberger et al., 2015).

Substituting isoleucine for leucine at position 6 results in a stabilization of the intrinsic repeat when compared to a consensus R-repeat. In crystal structures, leucine at this position is packed internally within the repeat (Figure 2.8B). Moving the leucine branched carbon from the γ-carbon to the β-carbon in isoleucine restricts the conformations the main chain carbon can adopt, possibly stabilizing the repeat itself. Isoleucine substitution at position 6 also destabilizes the interfaces between repeats, with the greatest destabilization between two substituted repeats. This R*-R* effect is non-additive, that is, it exceeds the destabilization of the singly-substituted R-R* and R*-R repeats.

Substituting phenylalanine for leucine at position 21 results in a destabilization of the intrinsic repeat compared to a consensus R-repeat. We also observe a modest effect on interfacial stability with adjacent consensus repeats. Leucine 21 forms a hydrophobic cluster in the core of the protein with leucine 6 and proline 5 within the same repeat, along with valine 28 of the previous repeat (Figure 2.8B). The L21F substitution likely disrupts this cluster with the introduction of the bulky phenyl group causing destabilization of the intrinsic repeat. When two repeats containing the same L21F substitution are adjacent to one another, the interface between the two is further destabilized to a greater

extent than when sharing an interface with a consensus repeat. The degree of destabilization is independent of N- and C-terminal orientation.

This same phenomenon is observed for the V28P substitution. Substituting proline for valine at position 28 increases the intrinsic folding stability, most likely due to the removal of the buried valine residue. This position is in a loop region of the repeat, and introduction of a proline residue decreases the accessible backbone degrees of freedom. Again, the greatest decrease in stability is observed when both substituted repeats are adjacent to one another. All four substitutions cause a significant decrease in repeat m-value, providing a hint at a decreased buried surface area when substitutions away from consensus are made.

2.3 Figures and Tables

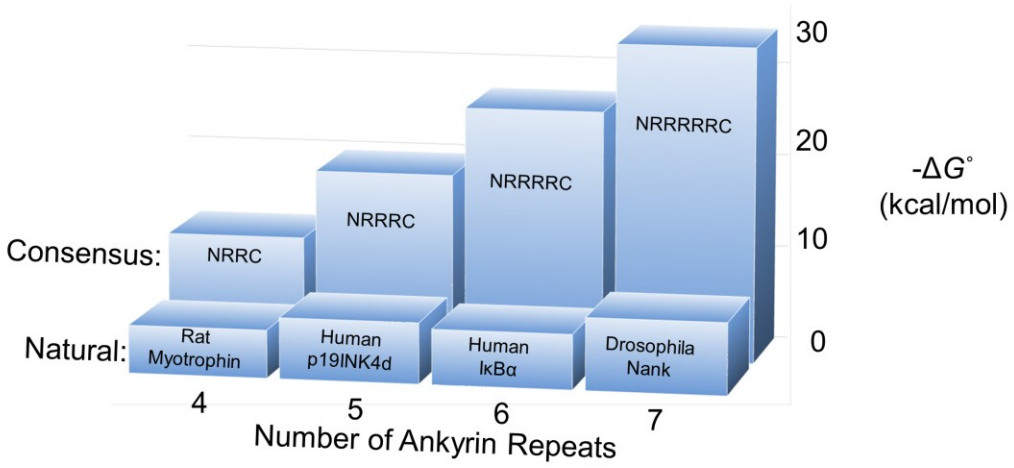


Figure 2.1 Comparison of folding free energies of naturally occurring ankyrin repeat proteins with consensus ankyrin repeat proteins of identical repeat number. Shown left to right: proteins with 4, 5, 6, and 7 ankyrin repeats occurring in nature (front) compared to consensus ankyrin repeats of identical repeat length (background). Free energy values from rat myotrophin (Mosavi et al., 2002), human p19INK4d (Löw et al., 2007), human IκBα (DeVries et al., 2011), and the ankyrin domain of the Drosophila Notch receptor (Tripp and Barrick, 2004). Consensus ankyrin repeat free energy values determined from Ising analysis (Aksel et al., 2011).

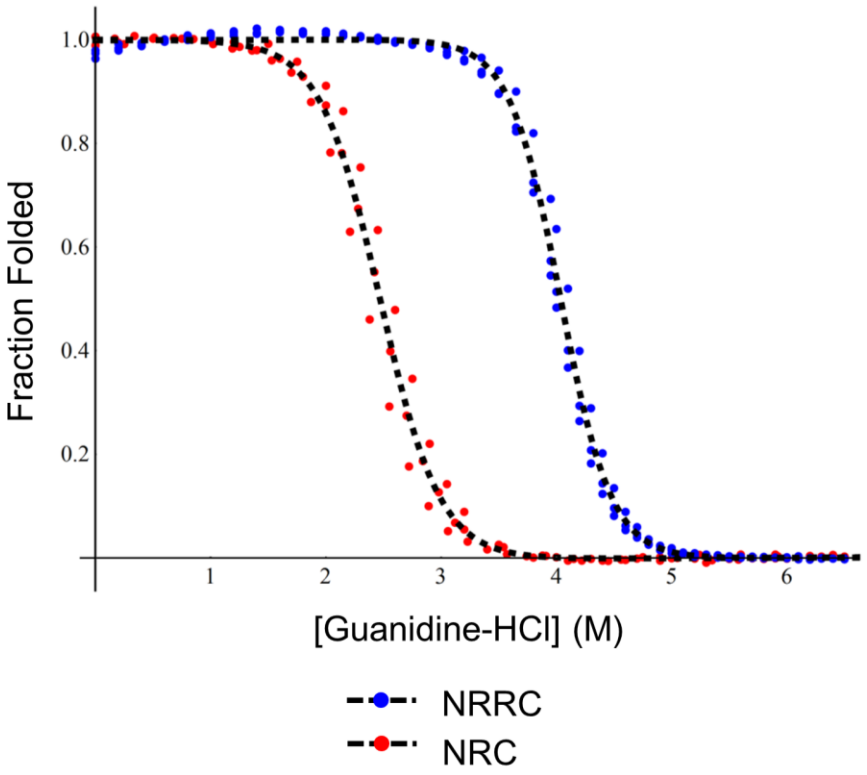


Figure 2.2 Guanidine-HCl unfolding transitions of 3-repeat and 4-repeat consensus ankyrin repeat proteins. Triplicate data are fit individually to obtain parameters using a two-state model. Shown are average fits (lines) to triplicate date (circles). Data and fits are normalized after fitting by subtracting the fitted baselines. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C.

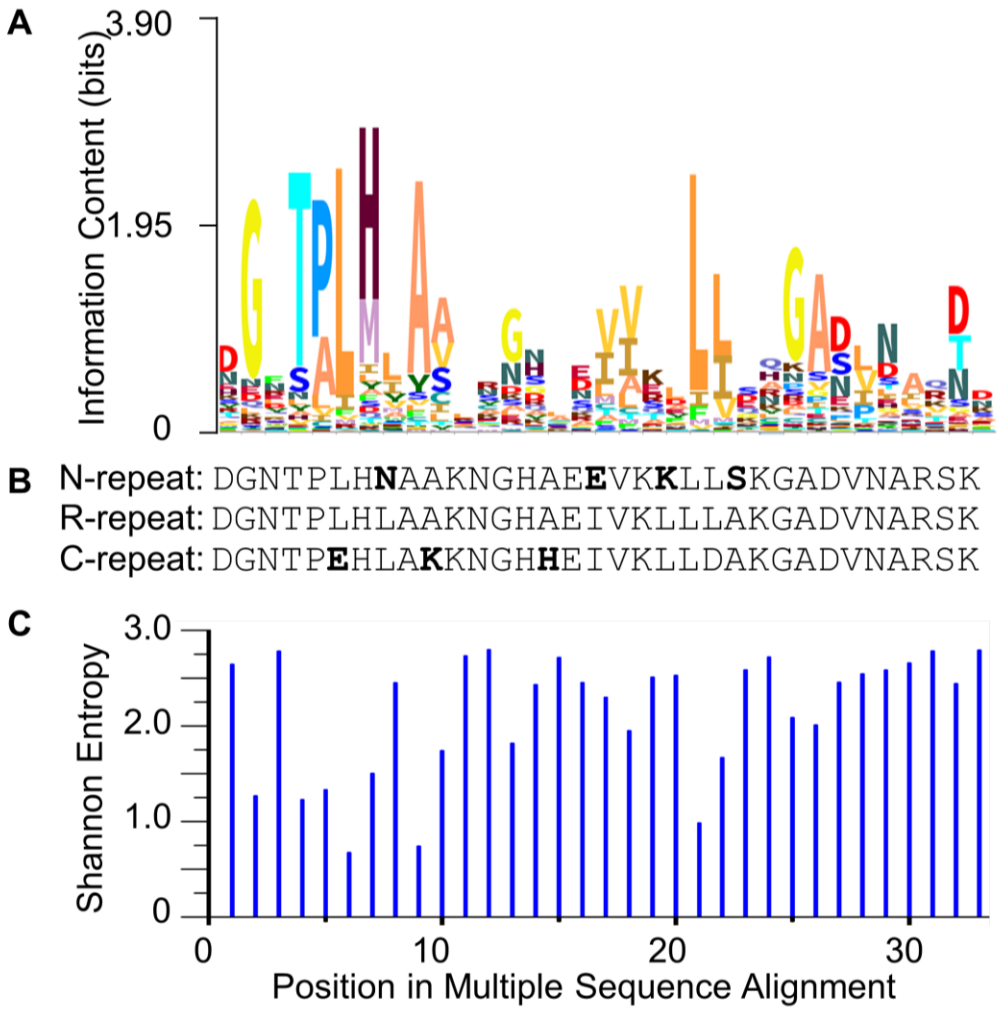


Figure 2.3 Sequence variation of ankyrin repeats derived from multiple sequence alignment. Multiple sequence seed alignment from Pfam database entry: Ank (PF00023). (A) Skylign Hidden Markov Model (HMM) logo for ankyrin the repeat (Wheeler et al., 2014). (B) Consensus ankyrin repeat sequence R (Wetzel et al., 2008), along with modified N- and C-capping repeat sequences (Aksel et al., 2011). Residues shown in bold are polar substitutions made for solubilizing N- and C-capping repeats. (C) Shannon Entropy calculated for Ank multiple sequence alignment (Entropy-One: www.hiv.lanl.gov)

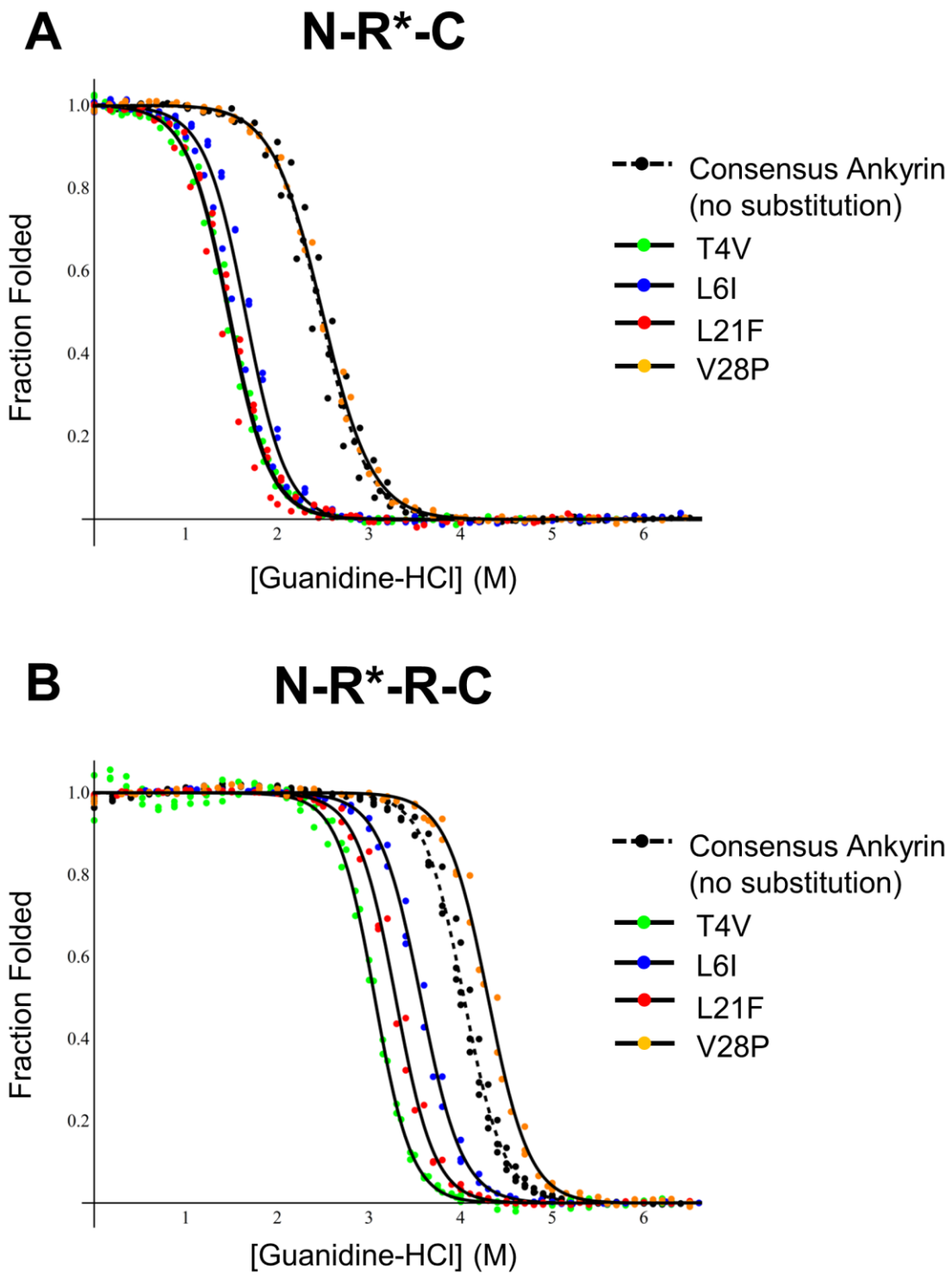
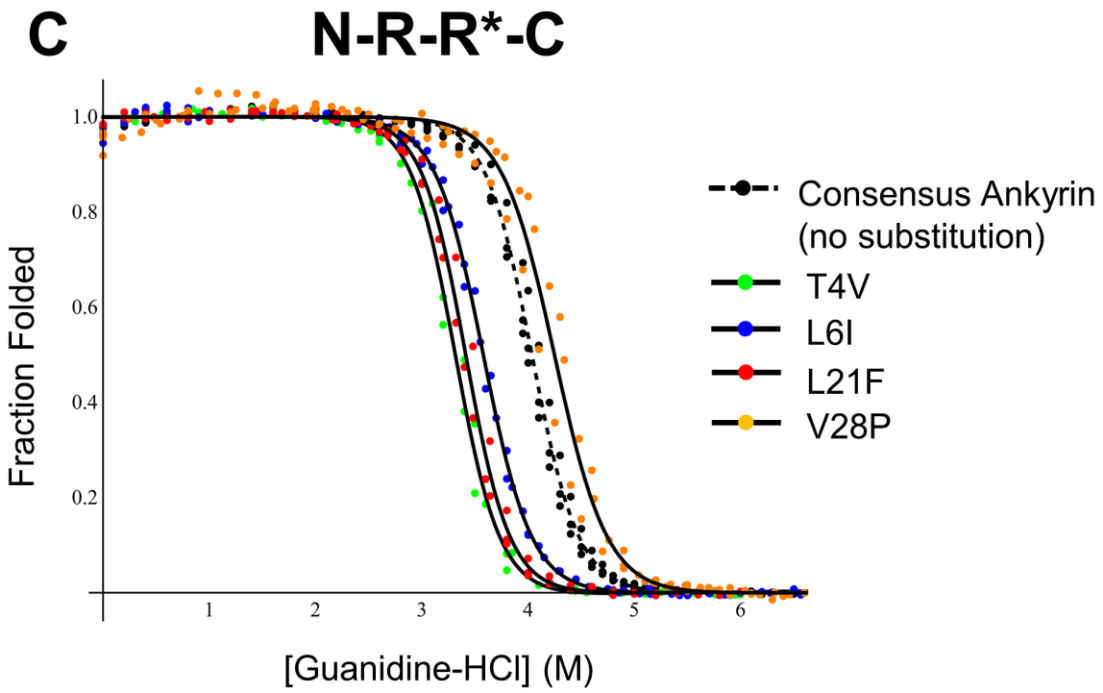
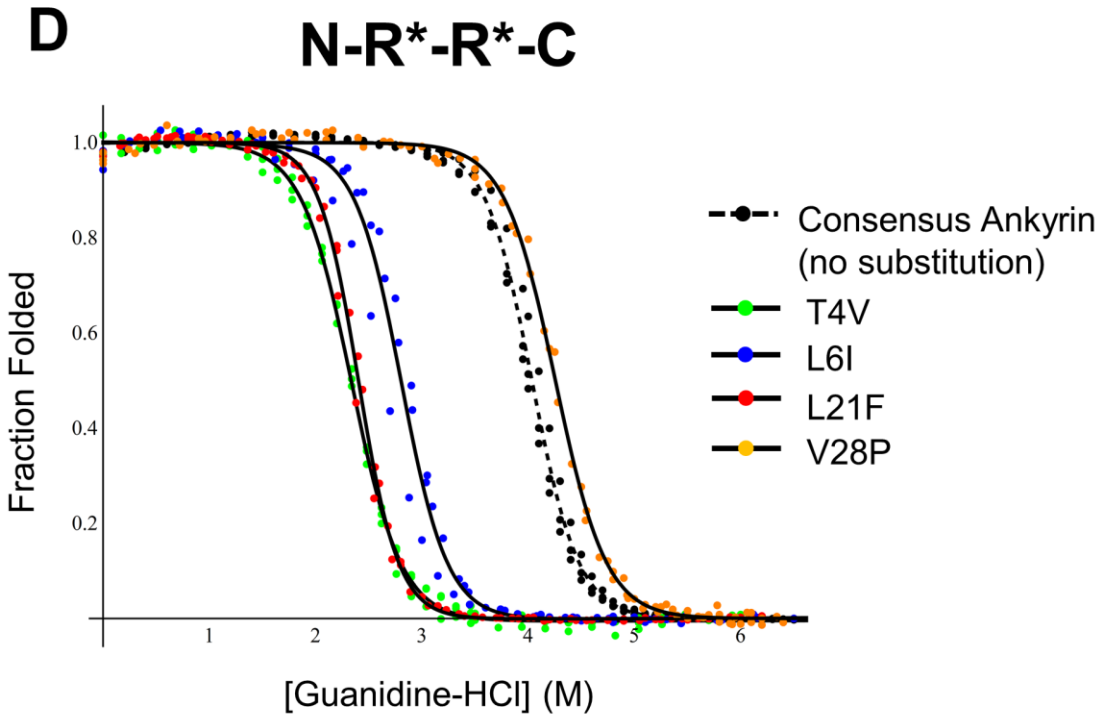


Figure 2.4 Guanidine-HCl unfolding transitions of three- and four-repeat consensus ankyrin constructs with point substitutions. Point substitutions are indicated in the legend. R* indicates the substituted repeat for (A) three-repeat NRC and (B,C,D) four-repeat NRRC. Consensus ankyrin repeats with no substitutions…(continued next page)





…(continued) are shown in black with dashed lines. Triplicate data are fit individually to obtain parameters using a two-state model. Shown are average fits to triplicate date (circles). Data and fits are normalized after fitting by subtracting the fitted baselines. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C.

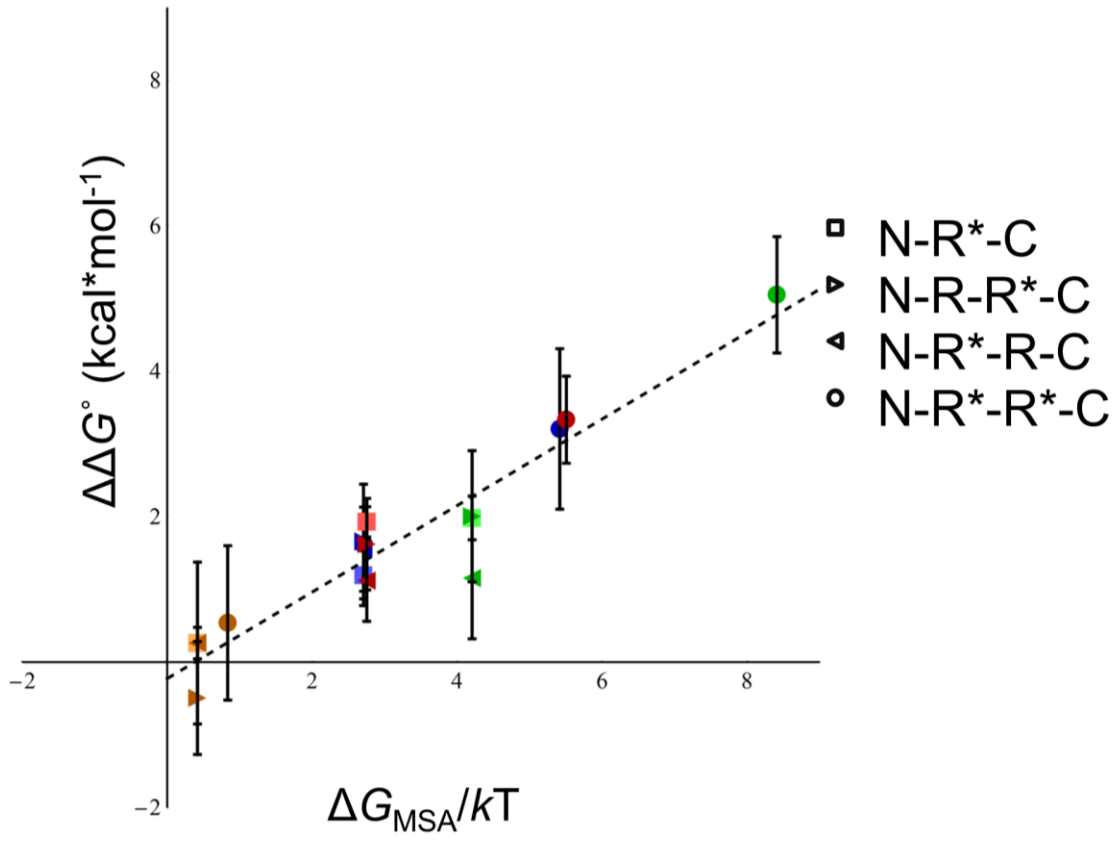


Figure 2.5 Substitutions away from consensus result in a proportional change to folding free energy. Changes in folding free energy are determined from two-state fit parameters for consensus ankyrin repeats. Error is determined by the square root of the sum of the squared errors for each independent value. Changes in multiple sequence alignment free energy are calculated by $-\ln(p^*/p^{wt})$ where p^* is the MSA probability of the substituting residue at the given position, and p^{wt} is the probability of the consensus residue. Key indicates the protein construct, and colors indicate substitutions as follows: T4V: Green, L6I: Blue, L21F: Red, V28P: Tan. Linear fit to data (slope: 0.59, correction coefficient: -0.22) shown with dashed line (p-value: $1.48112*10^{-8}$).

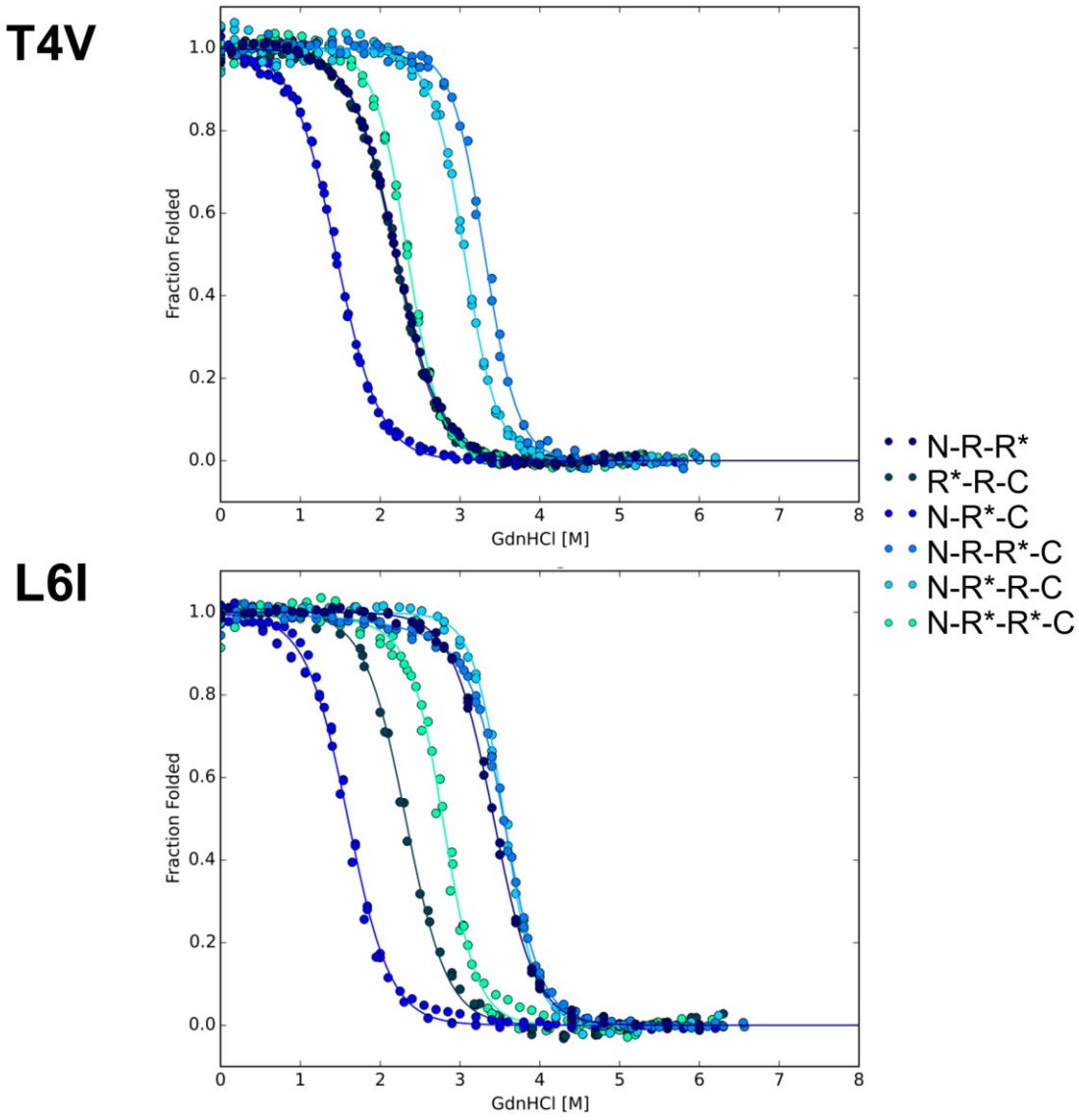


Figure 2.6 Ising model analysis of guanidine-HCl induced unfolding transitions of consensus ankyrin repeat protein variants. R* indicates the individual R repeats that contain the point substitutions. Data are globally fitted with a nearest-neighbor Ising model, and curves are normalized to fraction folded after fitting by subtracting the fitted baselines. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C. (continued on next page)

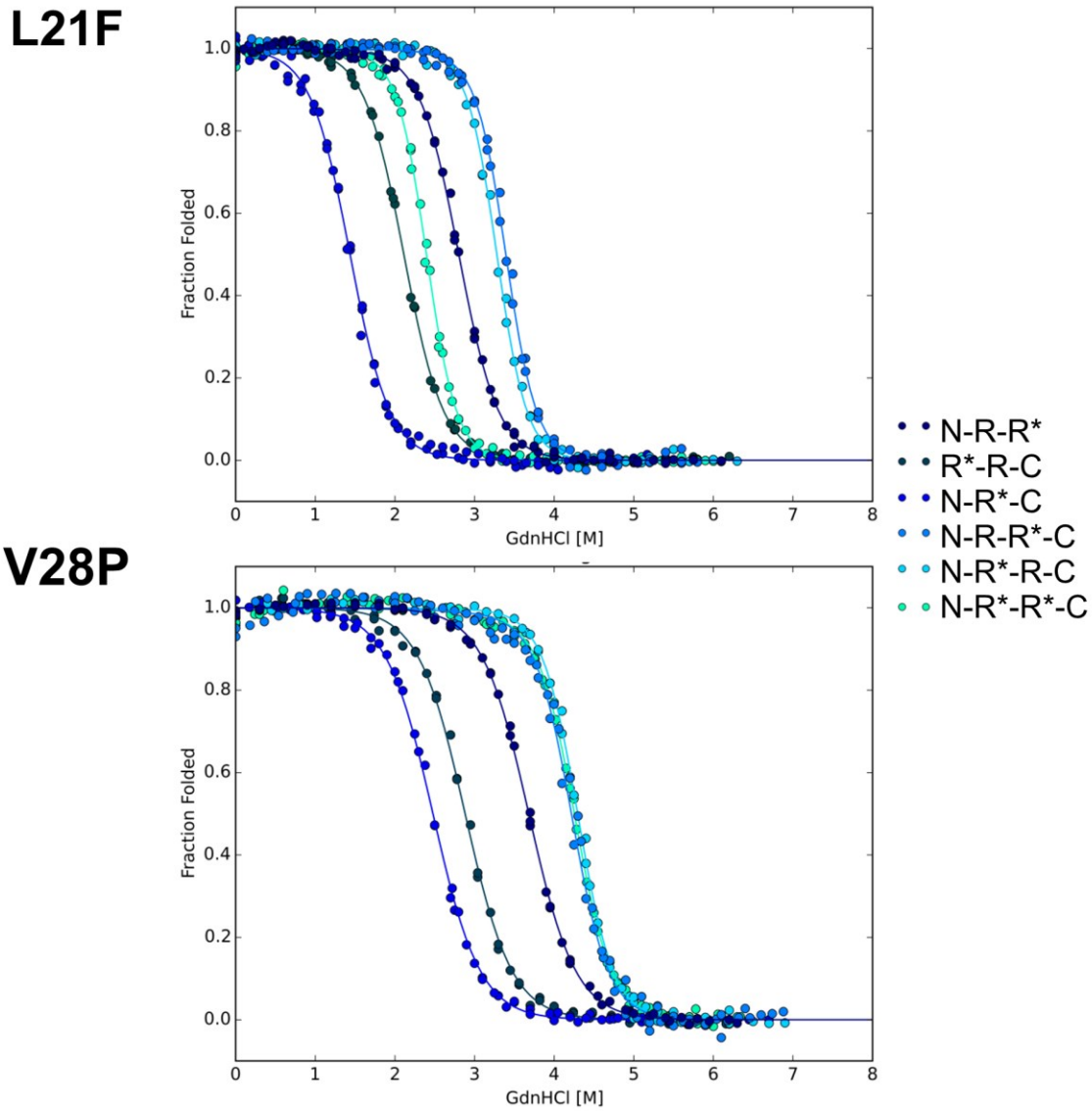


Figure 2.6 (continued) Ising model analysis of guanidineHCl-induced unfolding transitions of consensus ankyrin repeat protein variants. R* indicates the individual R repeats that contain the point substitutions. Data are globally fitted with a nearest-neighbor Ising model, and curves are normalized to fraction folded after fitting by subtracting the fitted baselines. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C.

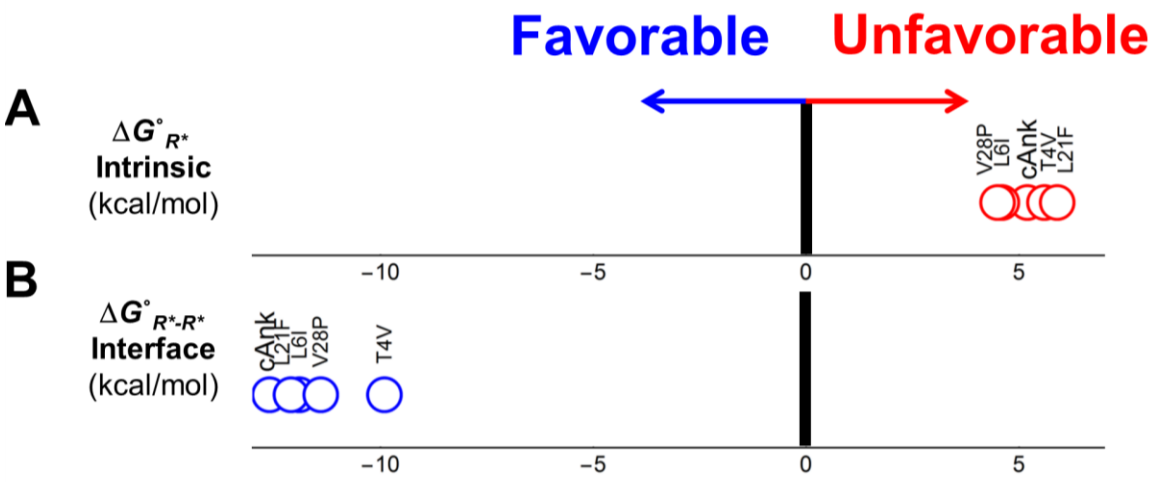


Figure 2.7 Non-consensus residue substitutions destabilize the highly favorable repeat interface. (A) Intrinsic and (B) interfacial coupling free energies determined by Ising analysis for substitutions (T4V, L6I, L21F, V28P) away from the consensus ankyrin "cAnk" sequence (Aksel et al., 2011). Non-consensus residues have variable effects on the unfavorable intrinsic repeat stability. However, the highly favorable consensus ankyrin repeat interface is destabilized by substitutions away from the consensus sequence.

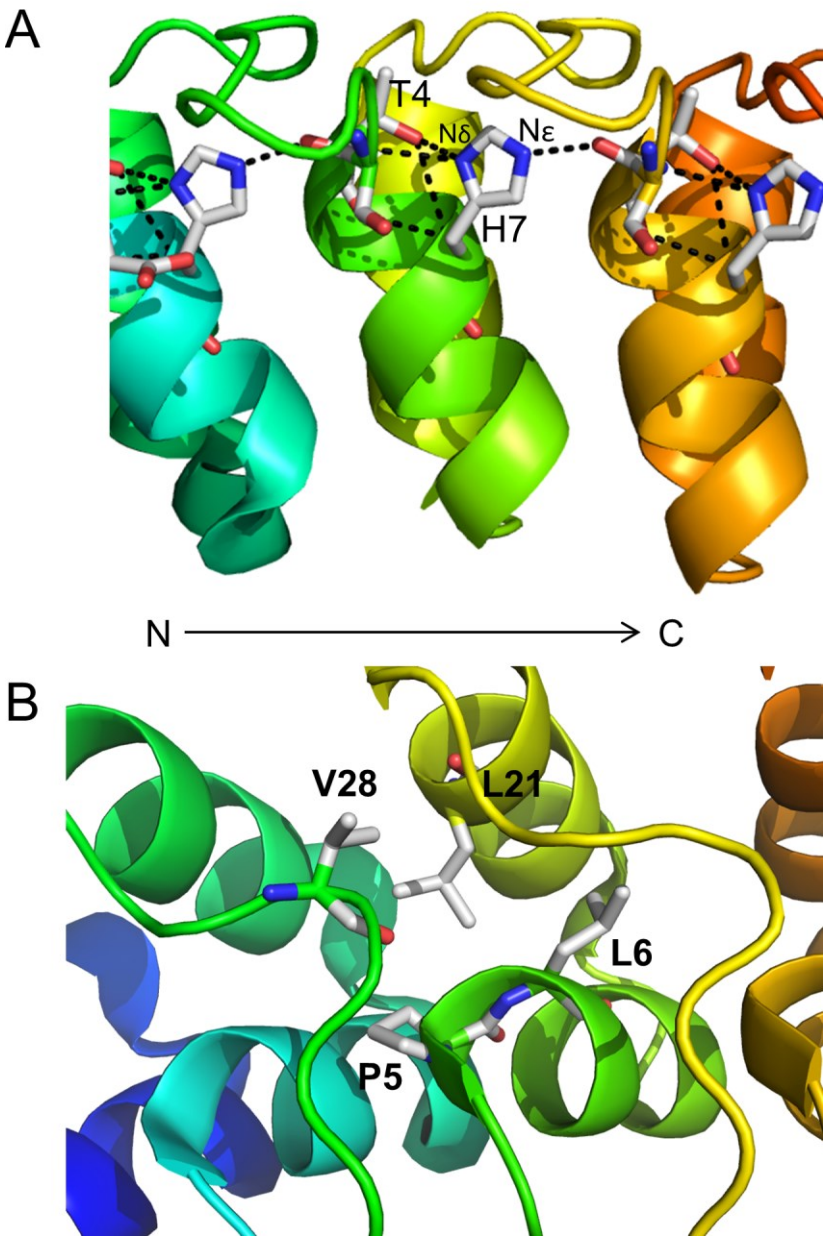


Figure 2.8 Crystal structure of ankyrin repeat substitution positions. PDB: 2BKG (A) Thr 4 plays an important role in a buried hydrogen bond ladder in ankyrin repeats. The Thr 4-OH and backbone NH in each repeat takes part in a bifurcated hydrogen bond to H7-Nδ within the same repeat. H7-Nε hydrogen bonds to backbone CO at position 3 of the next C-terminal repeat. (B) L6 and L21 form an extended hydrophobic core with V28 of the previous repeat in the core of ankyrin repeat array.

Table 2.1 Ankyrin Residue Sequence Entropy

Consensus Residue	**T4**	**L6**	**L21**	**V28**
Probability	0.737	0.854	0.831	0.220
Substituted Residue	**V**	**I**	**F**	**P**
Probability	0.011	0.053	0.057	0.145

Substitution	**T4V**	**L6I**	**L21F**	**V28P**
ΔG_{MSA}	4.20	2.78	2.68	0.42

Probabilities determined from seed multiple sequence alignment for Pfam (1063 sequences) Entry: Ank (PF00023)

Table 2.2 Parameters obtained from gdn-HCl titration melts at pH8, 20˚C

Repeat Protein:	**N-R*-C**		**N-R*-R-C**	
R* Repeat Substitution	-ΔG°	*m*-value	-ΔG°	*m*-value
T4V	3.8 ± 0.23	2.6 ± 0.08	8.8 ± 0.72	2.9 ± 0.21
L6I	4.6 ± 0.37	2.8 ± 0.09	9.2 ± 0.57	2.6 ± 0.13
L21F	3.8 ± 0.08	2.6 ± 0.19	9.2 ± 0.33	2.8 ± 0.04
V28P	5.5 ± 0.09	2.2 ± 0.05	11.3 ± 0.56	2.6 ± 0.08
Repeat Protein:	**N-R-R*-C**		**N-R*-R*-C**	
R* Repeat Substitution	-ΔG°	*m*-value	-ΔG°	*m*-value
T4V	9.7 ± 0.64	2.9 ± 0.18	5.8 ± 0.59	2.5 ± 0.25
L6I	9.3 ± 0.21	2.6 ± 0.02	7.6 ± 0.96	2.7 ± 0.22
L21F	9.7 ± 0.15	2.9 ± 0.03	7.5 ± 0.26	3.1 ± 0.05
V28P	10.6 ± 0.97	2.5 ± 0.16	10.3 ± 0.92	2.4 ± 0.20

Free energies and *m*-values are in kcal*mol^{-1} and kcal*mol^{-1}*M^{-1} respectively and are obtained from a two-state fit. Errors are determined from the standard deviation about the mean for three independent titrations.

Table 2.3 Ising parameters for consensus ankyrin amino acid substitutions

R* repeat substitution	Fit Parameters			
	$\Delta G^{\circ}_{R^*}$	$\Delta G^{\circ}_{N\text{-}R^*}$	$\Delta G^{\circ}_{R\text{-}R^*}$	*m*-value
T4V	5.6 ± 0.15	-10.2 ± 0.10	-10.7 ± 0.21	-0.7 ± 0.01
L6I	4.6 ± 0.17	-11.6 ± 0.16	-10.2 ± 0.28	-0.8 ± 0.02
L21F	5.9 ± 0.12	-11.4 ± 0.11	-11.3 ± 0.19	-0.8 ± 0.01
V28P	4.5 ± 0.16	-11.8 ± 0.17	-10.7 ± 0.28	-0.7 ± 0.02

R* repeat substitution	Fit Parameters		
	$\Delta G^{\circ}_{R^*\text{-}R}$	$\Delta G^{\circ}_{R^*\text{-}R^*}$	$\Delta G^{\circ}_{R^*\text{-}C}$
T4V	-9.2 ± 0.22	-9.2 ± 0.21	-12.5 ± 0.12
L6I	-11.1 ± 0.33	-9.3 ± 0.28	-10.5 ± 0.13
L21F	-11.3 ± 0.22	-10.0 ± 0.19	-11.7 ± 0.09
V28P	-10.7 ± 0.32	-10.8 ± 0.30	-11.6 ± 0.13

Free energies and m-values are in kcal*mol^{-1} and kcal*mol^{-1}*M^{-1} respectively and are obtained from a global Ising fit. Confidence intervals (at the 95% level) are obtained by bootstrap analysis (1000 iterations), assuming parameter estimates to be normally distributed. The intrinsic free energy of each substituted repeat is represented as $\Delta G^{\circ}_{R^*}$.The free energy of the interface between the repeats X and R* is represented as $\Delta G^{\circ}_{X\text{-}R^*}$, $\Delta G^{\circ}_{R^*\text{-}X}$, where X is repeat N, R, C, or R*. Denaturant effects are modeled with a single m-value (m_i).

2.4 Discussion

The motivation for this study is to better understand the relationship between conserved amino acids and the thermodynamic stability they confer to consensus proteins using consensus ankyrin repeat proteins. The large number of available protein sequences in the ankyrin multiple sequence alignment provides for a robust sampling of protein sequence space, and the resulting consensus designed protein has been shown to be more stable than natural ankyrin repeats (Tripp and Barrick, 2007). The simple linear, modular repeat helical architecture of ankyrin repeats has permitted the use of nearest-neighbor Ising analysis to determine the intrinsic and interfacial contributions to folding free energy (Aksel and Barrick, 2009; Aksel et al., 2011).

Crystal structures of ankyrin repeats have resolved interacting players in the protein core and along the surface (Binz et al., 2006; Mosavi et al., 2002); however to date it has been unclear what thermodynamic role the conserved amino acid residues play in stabilizing the ankyrin repeat protein fold. Information theory (Shannon, 1948) provides a framework to analyze large sets of aligned, related sequences. By determining the relative sequence entropy at each position of the

ankyrin multiple sequence alignment, we can obtain a numerical measure that numerical sequence variation has within the alignment, and compare to experimental thermodynamic measurements. Substitutions away from the consensus sequence carry a proportional change in contributions to the free energy within the distribution of amino acids at that residue position. By making different iterative protein construct arrays containing these substitutions, we have been able to determine the global effects these substitutions have on global folding free energy. The two-state cooperative folding behavior of ankyrin repeats allows for direct comparison between two-state global stabilities and representative multiple sequence alignment free energies associated with substitutions. We observe a linear relationship between the relative levels of sequence conservation of amino acid players in a substitution along with the changes in global folding free energy these substitutions cause in consensus background. From this correlation we conclude that the highly conserved consensus amino acids contribute more to the stability of the protein than less-conserved amino acids.

Furthermore, in the present study we employ a new variation of the nearest-neighbor Ising analysis to resolve the effects point substitutions away from the consensus sequence have on intrinsic folding and interfacial coupling free energies. Previous studies have dissected the nature of a buried threonine-histidine hydrogen-bonding network within the core of ankyrin repeat proteins (Preimesberger et al.,

2015). The current study confirms the thermodynamic role of Thr 4 in intrinsic stability and in particular, particular interfacial coupling free energies. Furthermore, substituting non-consensus residues into a buried hydrophobic pocket has allowed us to tease out the energetic consequences of shuffling carbon-carbon bonds in a hydrophobic core. Further analysis by solution NMR is required to deduce the true changed environment these substitutions cause.

The studies presented here have clear implications for information theory as it relates to consensus protein design and protein folding thermodynamics. To understand the effects all positions have in the ankyrin repeat sequence, a better more comprehensive study is required to fully sample sequence space across each of the 33 residues. Future studies are needed to investigate the effects that conservative and non-conservative substitutions have on consensus ankyrin repeats. What regions of the consensus ankyrin repeat sequence can tolerate more substitutions than others? How does sequence conservation for surface residues along binding interfaces compare to other regions of the protein? Do levels of sequence conservation scale differently with repeat protein length?

2.5 Materials and Methods

2.5.1 Cloning, expression, and purification

To clone arrays of consensus ankyrin repeat variants, we employed a complementary BamHI/BglII ligation strategy in a modified pET15b expression vector (Novagen) as described previously (Aksel et al., 2011). All constructs were confirmed by DNA sequencing.

E. coli BL21(DE3) were transformed with plasmids containing repeat protein genes, and were grown in Luria Broth at 37˚C to an OD_{600} of 0.6-0.8. Expression was induced by addition of 1 mM IPTG. After further growth for 4-6 hours at 37˚C, cells were collected by centrifugation and frozen at -80˚C. Cell pellets were resuspended in 8 M urea, 1 M NaCl, and 25 mM Tris-HCl pH 8.0, lysed by sonication, and centrifuged to remove insoluble cell debris. The supernatant was loaded onto a nickel column (QIAGEN). After washing with the same resuspension buffer, bound protein was refolded on the column by washing with 150 mM NaCl, 25 mM Tris-HCl pH 8.0, and subsequently eluted with 0.5 M imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 8.0.

Pure protein-containing fractions were dialyzed overnight into the desired buffer to remove imidazole.

2.5.2 Multiple sequence alignment and sequence entropy calculations

The Pfam seed ankyrin multiple sequence alignment (Pfam entry: Ank PF00023) was used for all probability calculations. Sequence entropy calculations and sequence entropy plots were generated using the HIV sequence database online Entropy-One calculation tool (www.hiv.lanl.gov). Ankyrin HMM logos were designed using Skylign HMM logo online (http://skylign.org/; see (Wheeler et al., 2014)).

2.5.3 Circular dichroism spectroscopy and guanidine-HCl induced unfolding transitions

All CD measurements were done using AVIV Model 400 CD spectrometers (AVIV Associates, Lakewood, NJ). Guanidine-HCl induced

unfolding transitions were obtained using a Hamilton 500 titrator, and were monitored by CD at 222 nm. Protein concentrations ranged from 2-4 μM. Measurements were made in a silanized 1 cm quartz cuvette. Signal was averaged for 30 seconds at each denaturant concentration, and a delay of 5 minutes, several times the signal averaging time, was used to ensure complete mixing and equilibration prior to data acquisition.

Two-state fits to denaturant induced unfolding transitions were performed using a two-state unfolding model (Greene and Pace, 1974) with Wolfram Mathematica fitting software.

Intrinsic and interfacial free energies were determined by global fitting with a heteropolymer Ising model as previously described (Geiger-Schuller and Barrick, 2016) adapted to resolve substituted R* repeats. In this model, intrinsic and interfacial folding free energies are represented using equilibrium constants κ and τ respectively (Equations 2.7-2.16).

$$\kappa_N = e^{-(\Delta G_N - \mathrm{m}[GdnHCl])/RT} \quad (2.7)$$

$$\kappa_R = e^{-(\Delta G_R - m[GdnHCl])/RT} \quad (2.8)$$

$$\kappa_C = e^{-(\Delta G_C - m[GdnHCl])/RT} \quad (2.9)$$

$$\kappa_{R*} = e^{-(\Delta G_{R*} - m[GdnHCl])/RT} \quad (2.10)$$

$$\tau_{N-R} = \tau_{R-C} = e^{-(\Delta G_{N-R})/RT} \quad (2.11)$$

$$\tau_{N-R*} = e^{-(\Delta G_{N-R})/RT} \quad (2.12)$$

$$\tau_{R-R*} = e^{-(\Delta G_{R-R*})/RT} \quad (2.13)$$

$$\tau_{R*-R} = e^{-(\Delta G_{R*-R})/R\mathrm{T}} \quad (2.14)$$

$$\tau_{R*-R*} = e^{-(\Delta G_{R*-R*})/RT} \quad (2.15)$$

$$\tau_{R*-C} = e^{-(\Delta G_{R*-C})/RT} \quad (2.16)$$

In this analysis, the intrinsic free energies of the N, R, C, and R* repeats are treated as separate parameters (Equations 2.7-2.11). The value of the intrinsic energy of folding of R* is adjusted in our fits. Values for the intrinsic folding of the N, R, and C repeats have been previously determined (Aksel et al., 2011) and were treated as constants in our fits, as were the interfacial free energies of N-R and R-C (Equation 2.11). The interfacial free energies of N-R*, R-R*, R*-R, R*-R*, and R*-C, (Equations 2.12-2.16) were treated as separate adjustable parameters. Using these equilibrium constants, the partition function, q, can be constructed for each repeat construct by multiplying two-by-two transfer matrices (Equations 2.17-22). In this analysis, each matrix associates the κ term with the τ of the next repeat.

$$q_{NR*C} = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\kappa_N\tau_{N-R*} & \kappa_N \\ 1 & 1\end{bmatrix}\begin{bmatrix}\kappa_{R*}\tau_{R*-C} & \kappa_{R*} \\ 1 & 1\end{bmatrix}\begin{bmatrix}\kappa_C\tau_{N-R} & \kappa_C \\ 1 & 1\end{bmatrix}\begin{bmatrix}0 \\ 1\end{bmatrix} \quad (2.17)$$

$$q_{NRR*} = \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}\kappa_N\tau_{N-R} & \kappa_N \\ 1 & 1\end{bmatrix}\begin{bmatrix}\kappa_R\tau_{R-R*} & \kappa_R \\ 1 & 1\end{bmatrix}\begin{bmatrix}\kappa_{R*}\tau_{R*-R} & \kappa_{R*} \\ 1 & 1\end{bmatrix}\begin{bmatrix}0 \\ 1\end{bmatrix} \quad (2.18)$$

$$q_{R*RC} = [1 \quad 1] \begin{bmatrix} \kappa_{R*}\tau_{R*-R} & \kappa_{R*} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R\tau_{R-C} & \kappa_R \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_C\tau_{N-R} & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.19}$$

$$q_{NRR*C} = [1 \quad 1] \begin{bmatrix} \kappa_N\tau_{N-R} & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R\tau_{R-R*} & \kappa_R \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_{R*}\tau_{R*-C} & \kappa_{R*} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_C\tau_{N-R} & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.20}$$

$$q_{NR*RC} = [1 \quad 1] \begin{bmatrix} \kappa_N\tau_{N-R*} & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_{R*}\tau_{R*-R} & \kappa_{R*} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R\tau_{R-C} & \kappa_R \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_C\tau_{N-R} & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.21}$$

$$q_{NR*R*C} = [1 \quad 1] \begin{bmatrix} \kappa_N\tau_{N-R*} & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_{R*}\tau_{R*-R*} & \kappa_{R*} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_{R*}\tau_{R*-C} & \kappa_{R*} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_C\tau_{N-R} & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{2.22}$$

Ising parameters were determined by nonlinear least squares using an in-house Python program (written by J. Marold, adapted by K. Geiger-Schuller). Confidence intervals (at the 95% level) were obtained by bootstrap analysis (1000 iterations), assuming parameter estimates to be normally distributed.

2.6 References

Aksel, T., and Barrick, D. (2009). Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. Methods Enzymol. 455, 95–125.

Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993 19, 349–360.

Binz, H.K., Stumpp, M.T., Forrer, P., Amstutz, P., and Plückthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. J. Mol. Biol. 332, 489–503.

Binz, H.K., Kohl, A., Plückthun, A., and Grütter, M.G. (2006). Crystal structure of a consensus-designed ankyrin repeat protein: Implications for stability. Proteins Struct. Funct. Bioinforma. 65, 280–284.

DeVries, I., Ferreiro, D.U., Sánchez, I.E., and Komives, E.A. (2011). Folding kinetics of the cooperatively folded subdomain of the IκBα ankyrin repeat domain. J. Mol. Biol. 408, 163–176.

Geiger-Schuller, K., and Barrick, D. (2016). Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. Biophys. J. 111, 2395–2403.

Greene, R.F., and Pace, C.N. (1974). Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. J. Biol. Chem. 249, 5388–5393.

Kajander, T., Cortajarena, A.L., and Regan, L. (2006). Consensus Design as a Tool for Engineering Repeat Proteins. In Protein Design, (Humana Press), pp. 151–170.

Komor, R.S., Romero, P.A., Xie, C.B., and Arnold, F.H. (2012). Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Eng. Des. Sel. PEDS 25, 827–833.

Löw, C., Weininger, U., Zeeb, M., Zhang, W., Laue, E.D., Schmid, F.X., and Balbach, J. (2007). Folding mechanism of an ankyrin repeat protein: scaffold and active site formation of human CDK inhibitor p19(INK4d). J. Mol. Biol. 373, 219–231.

Mosavi, L.K., Minor, D.L., and Peng, Z. (2002). Consensus-derived structural determinants of the ankyrin repeat motif. Proc. Natl. Acad. Sci. 99, 16029–16034.

Mosavi, L.K., Cammett, T.J., Desrosiers, D.C., and Peng, Z.-Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. Protein Sci. Publ. Protein Soc. 13, 1435–1448.

Pascual, J., Castresana, J., and Saraste, M. (1997). Evolution of the spectrin repeat. BioEssays News Rev. Mol. Cell. Dev. Biol. 19, 811–817.

Preimesberger, M.R., Majumdar, A., Aksel, T., Sforza, K., Lectka, T., Barrick, D., and Lecomte, J.T.J. (2015). Direct NMR detection of bifurcated hydrogen bonding in the α-helix N-caps of ankyrin repeat proteins. J. Am. Chem. Soc. 137, 1008–1011.

Schüler, A., and Bornberg-Bauer, E. (2016). Evolution of Protein Domain Repeats in Metazoa. Mol. Biol. Evol. 33, 3170–3182.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423.

Tripp, K.W., and Barrick, D. (2004). The Tolerance of a Modular Protein to Duplication and Deletion of Internal Repeats. J. Mol. Biol. 344, 169–178.

Tripp, K.W., and Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. J. Mol. Biol. 365, 1187–1200.

Tripp, K.W., and Barrick, D. (2008). Rerouting the Folding Pathway of the Notch Ankyrin Domain by Reshaping the Energy Landscape. J. Am. Chem. Soc. 130, 5681–5688.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. 376, 241–257.

Wheeler, T.J., Clements, J., and Finn, R.D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics 15, 7.

Zweifel, M.E., Leahy, D.J., Hughson, F.M., and Barrick, D. (2003). Structure and stability of the ankyrin domain of the Drosophila Notch receptor. Protein Sci. Publ. Protein Soc. 12, 2622–2632.

Chapter 3

The unusual stability distributions of *de novo* designed helical repeat arrays: extreme global stability is determined by short-range interactions

This chapter includes contributions from Katie Geiger-Schuller and Max Yuhas in the Barrick lab, and is a collaboration with David Baker and Fabio Parmegianni at the University of Washington.

3.1 Introduction

Linear repeat proteins have proven to be useful model systems in the quest to better understand protein folding thermodynamics. Due to their

repetitive primary structures, these proteins fold into linearly extended modular arrays with translational symmetry. Unlike globular proteins, where interactions can span across the protein sequence, the interactions of linear repeat proteins are confined to within or between adjacent repeats (Kloss et al., 2008). This architecture simplifies the models used to describe protein folding thermodynamics, permitting the use of nearest-neighbor Ising analysis.

One-dimensional Ising analysis has been successfully applied to a number of linear helical repeat proteins (Aksel et al., 2011; Geiger-Schuller and Barrick, 2016; Kajander et al., 2005; Marold et al., 2015). This model assumes that repeat protein stability can be parsed into intrinsic folding energies of individual repeats and coupling energies at the interfaces between adjacent folded repeats. Previous work characterizing linear repeat proteins derived from naturally occurring folds shows that individual repeats are unstable. In these proteins, stability (and cooperativity) originates in the favorable interfaces between adjacent repeats.

The use of linear repeat proteins as molecular scaffold and recognition partners has been exploited in a number of engineering applications. Consensus ankyrin repeats have enhanced the activity of engineered cellulases (Cunha et al., 2016), transcription activator-like effector proteins (TALEs) have been engineered for in genome editing (Christian et al., 2010; Li et al., 2011), and molecular chaperones have been fused to tetratricopeptide repeat proteins (TPRs) to increase substrate affinity (Cortajarena et al., 2004). While such linear repeat proteins designed from naturally occurring families have remarkable utility, the demand for novel molecular scaffolds in biomolecular engineering

extends beyond extant protein families. To meet this demand, the Baker lab has developed a series of designed helical repeat proteins (DHRs) with native-state architectures that extend beyond those of naturally occurring repeat proteins (Brunette et al., 2015).

Here we characterize the stability of a series of DHRs using nearest-neighbor Ising analysis. We find that unlike naturally occurring repeat proteins, both the intrinsic and interfacial contributions to folding free energy of DHRs are thermodynamically favorable, giving rise to extraordinarily high folding stability while maintaining cooperativity. The favorable local stability of DHR repeats suggests a reduced kinetic folding barrier; we present kinetic measurements that confirm this prediction.

3.2 Results

3.2.1 Folding behavior of Designed Helical Repeats

To investigate the thermodynamic folding behavior of Rosetta-designed repeat proteins with novel fold geometries, we chose DHR candidates for characterization based on the following criteria: (1) available SAXS and crystal structure data that conform the original design criteria, (2) an absence of cysteine residues to reduce complications associated with disulfide linkages, and (3) experimental evidence that shows the capped repeat proteins to be monomeric in solution. The proteins DHR9, DHR10, DHR54, DHR71, and DHR79 satisfy these criteria. While DHRs with both caps are monomeric, we found they are prone to forming soluble aggregates when one cap is removed. For all DHRs studied here, addition of 10% glycerol prevented self-association in DHR constructs with either the N- and C- terminal caps and both N- and C-terminal caps.

To confirm that DHRs maintain α-helical secondary structure, we collected far-UV CD spectra for each four repeat NR_2C DHR (Figure 1B). The spectra reveal characteristic minima at 208 nm and 222 nm, consistent with folded α-helical proteins. To measure DHR stability, we monitored guanidine-

HCl induced unfolding transitions using CD spectroscopy at 222nm. For DHR10, DHR54, DHR71, and DHR79, NR_2C constructs displayed a single sigmoidal unfolding transition which is fitted well with a two-state model (Figure 1C). DHR9 did not unfold across a range of temperatures, pH, and denaturant concentrations (data not shown), preventing further study from solution thermodynamics. The unfolding transitions of DHRs 54, 71, and 79 have high slopes and midpoints for unfolding. The unfolding transition of DHR10.2 occurs over a broader range of denaturant concentration with a low midpoint compared to the other DHRs. Especially for DHRs 54, 71, and 79, the steep guanidine unfolding transitions suggest a high level of cooperativity.

3.2.2 Length and capping dependence on stability

To determine the effects of variation in repeat number and the sequence substitutions associated with the N- and C-terminal capping repeats on stability, we constructed a series of constructs that delete terminal and internal repeats For DHR10, deletion of the C-terminal repeat leads to formation of soluble aggregates in solution. To prevent aggregation, we made a series of charged substitutions to solvent-exposed hydrophobic residues in the N-terminal capping repeat (V12K, I14E, V16E, L39R) and refer to this construct as DHR10.2.

For each of the four DHR constructs that displayed folding transitions, we measured unfolding curves for constructs with two, three, and four repeats. Two repeat constructs contain a single R repeat with either an N-terminal capping repeat (NR) or a C-terminal capping repeat (RC). Three repeat constructs contain one construct with a single R repeat with both N- and C-terminal capping repeats (NRC), or two R repeats with either an N- (NR_2) or C-terminal (R_2C) capping repeat. The four repeat construct contains two R repeats with both N- and C-terminal capping repeats (NR_2C). For DHR54 we were also able to construct and characterize a folded, stable single N repeat.

Stabilities of length and capping variants were monitored by guanidine-HCl induced unfolding transitions by CD spectroscopy at 222nm as described above (Figure 2). For all DHR proteins, stability increases as the number of repeats increases (compare DHR54 N to NR, DHRs 10.2, 71, 79 NR to NR_2, and all DHRs NRC to NR_2C). However, the capping repeats are generally less stabilizing than internal "R" repeats. Adding a C-terminal capping repeat to DHR10.2 NR increases the transition slope and midpoint, whereas adding a C-terminal capping to NR_2 increases the slope more than midpoint (compare NR_2 to NR_2C). The C-terminal capping repeat gives rise to a larger slope and midpoint than the N-terminal capping repeat (compare NR_2 to R_2C), suggesting greater intrinsic stability for the C-cap, or a more stabilizing R:C interface.

For DHR54 and DHR71, the unfolding midpoint for N-terminal capped R-repeats are higher than those with only a C-terminal capping repeat (compare NR to RC). While for DHR54 capping identity does not affect transition slope, adding a C-terminal capping repeat to DHR71 appears to result in multistate

unfolding behavior (compare NR to NRC, NR_2 to NR_2C). Moreover, the N-cap repeat shifts the unfolding transition of DHR54 to higher guanidine concentration (compare NR to RC).

3.2.3 Ising analysis extracts intrinsic and interfacial folding free energies for all DHRs in the absence of glycerol

To separate intrinsic and interfacial folding energies, a 1-D Ising model allowing for contributions to stability from both guanidine-HCl and glycerol was fitted to DHR guanidine-induced unfolding transitions collected at several glycerol concentrations (Aksel et al., 2011; Kajander et al., 2005; Wetzel et al., 2008). In this model, individual repeats are monitored as either folded or unfolded states. Thus, for an n-repeat array, there are 2^n configurations treated by the model. The energy of each configuration is determined by the intrinsic folding energy of each repeat as well as the coupling ("interfacial") free energies between consecutive repeats. By varying the repeat protein length and capping-repeat identity, we were able to extract the intrinsic (ΔG_i) and interfacial ($\Delta G_{i,i+1}$) free energies. By varying the concentrations of guanidine-HCl and glycerol, we were able to determine the dependence of intrinsic energies on guanidine and glycerol concentration ($m_{Gdn\text{-}HCl}$ and $m_{Glycerol}$). We assume that N-

cap, central, and C-cap repeats have identical m-values, except for DHR71 which requires a separate $m_{Gdn\text{-}HCl}$ for the C-cap repeat.

Because the sequences of the N- and C-terminal capping repeats differ from the sequence of central repeats, three intrinsic energies are included in the model (ΔG_N, ΔG_R, and ΔG_C). For all DHRs except DHR54, the model includes only one interfacial free energy ($\Delta G_{i,i+1}$). Although it is possible that the free energies between central repeats and capping repeats are not identical, it is not possible to extract parameters unless the unfolding energy of the lone cap can be measured. Because an unfolding transition of a lone N-cap repeat for DHR54 is observed, a more complicated model including a separate term for the interfacial energy between an N-cap repeat and the adjacent central repeat ($\Delta G_{N,i+1}$) can be invoked.

Figure 2 shows four global fits of the Ising model to DHR unfolding transitions. Although there are only six global thermodynamic parameters for the fits in Figure 2A and 2D and seven global thermodynamic parameters in Figures 3.2C and D, the model also includes separate baseline parameters for each unfolding transition (Table 3.1).

Unlike all previously measured intrinsic energies for naturally-derived consensus repeat proteins (Aksel et al., 2011; Wetzel et al., 2008; Kajander et al., 2005; Marold et al., 2015; Geiger-Schuller and Barrick, 2016), fitted DHR intrinsic folding energies are intrinsically favorable (Figure 3.3A). The interfacial free energies of central DHR repeats are all stabilizing, and are within the range of interfacial values for naturally-derived linear repeats (Figure 3.3B). Given the magnitude of the stabilizing interfacial free energies are equal

between the unnatural DHRs and natural repeat proteins (although greater for consensus ankyrin), the favorable intrinsic stabilities of the DHRs result in an even more favorable ΔG°_{Net} for each added repeat (Figure 3.3C). Thus, the exceptional stability of DHRs, and the high level of success of this design approach, results from very high quality design at the local (intra-repeat) level. DHR71 and DHR10.2 have capping intrinsic energies that are unfavorable, and this results in multi-state transitions seen in panel 2A and 2C. The squared-sum of the residuals was greatly decreased by including a separate, less cooperative, m-value for the C-terminal capping repeat of DHR71. For all DHRs, guanidine-HCl is destabilizing and glycerol is stabilizing, although the extent of glycerol stabilization varies among DHRs.

3.3 Figures and Tables

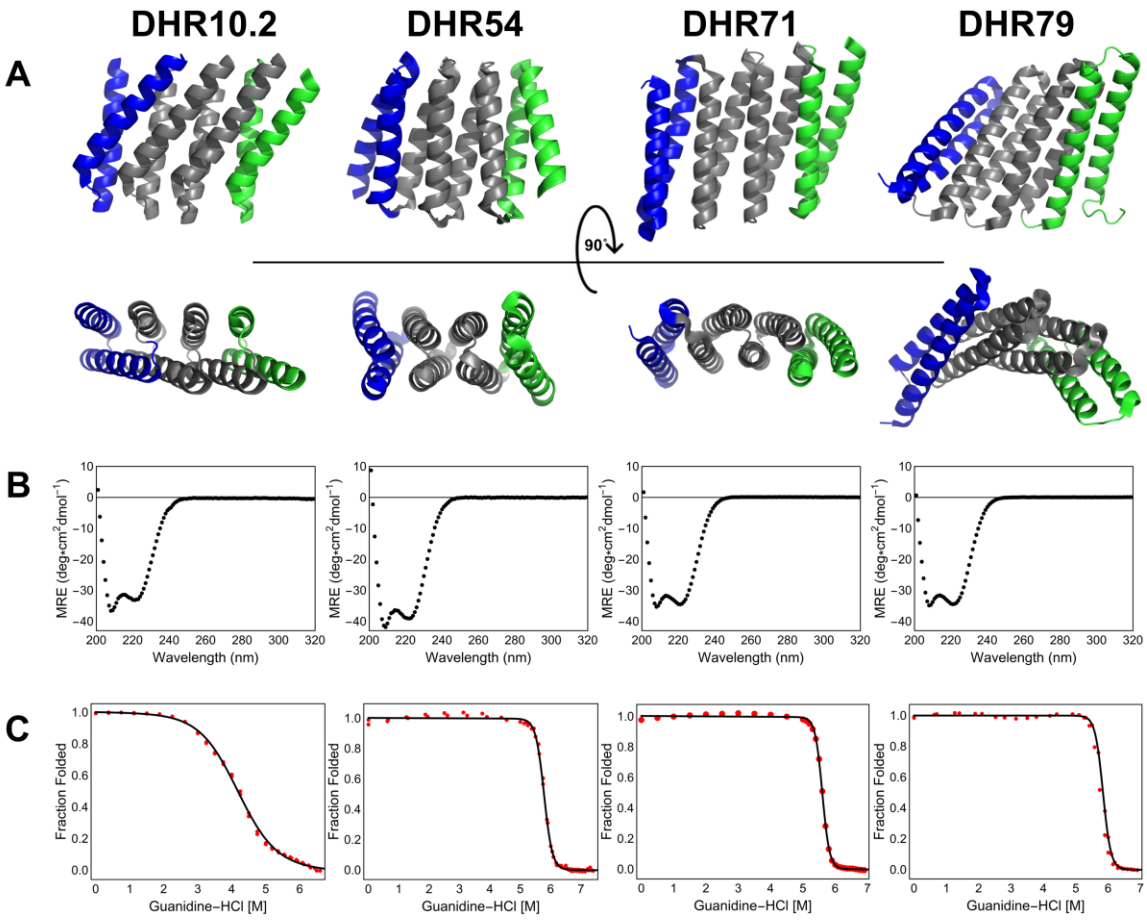


Figure 3.1 Structures and stabilities of designed helical repeat proteins. (A) Selected DHR proteins have distinct structures that are to-date unobserved in natural repeat proteins, including unique inter-repeat twists and radii of curvature between repeating units. (B) Far-UV circular dichroism shows characteristic α-helical spectra for DHR proteins. (C) Guanidine-induced unfolding transitions of four-repeat NR_2C DHR proteins (red circles) fit with a two-state unfolding model (black curves) reveal stable, cooperative folding behavior of the DHR proteins. Panels in (B) and (C) correspond to the DHR proteins shown in (A). PDB codes are 5CWG (DHR10), 5CWL (DHR54), 5CWN (DHR71), and 5CWP (DHR79).

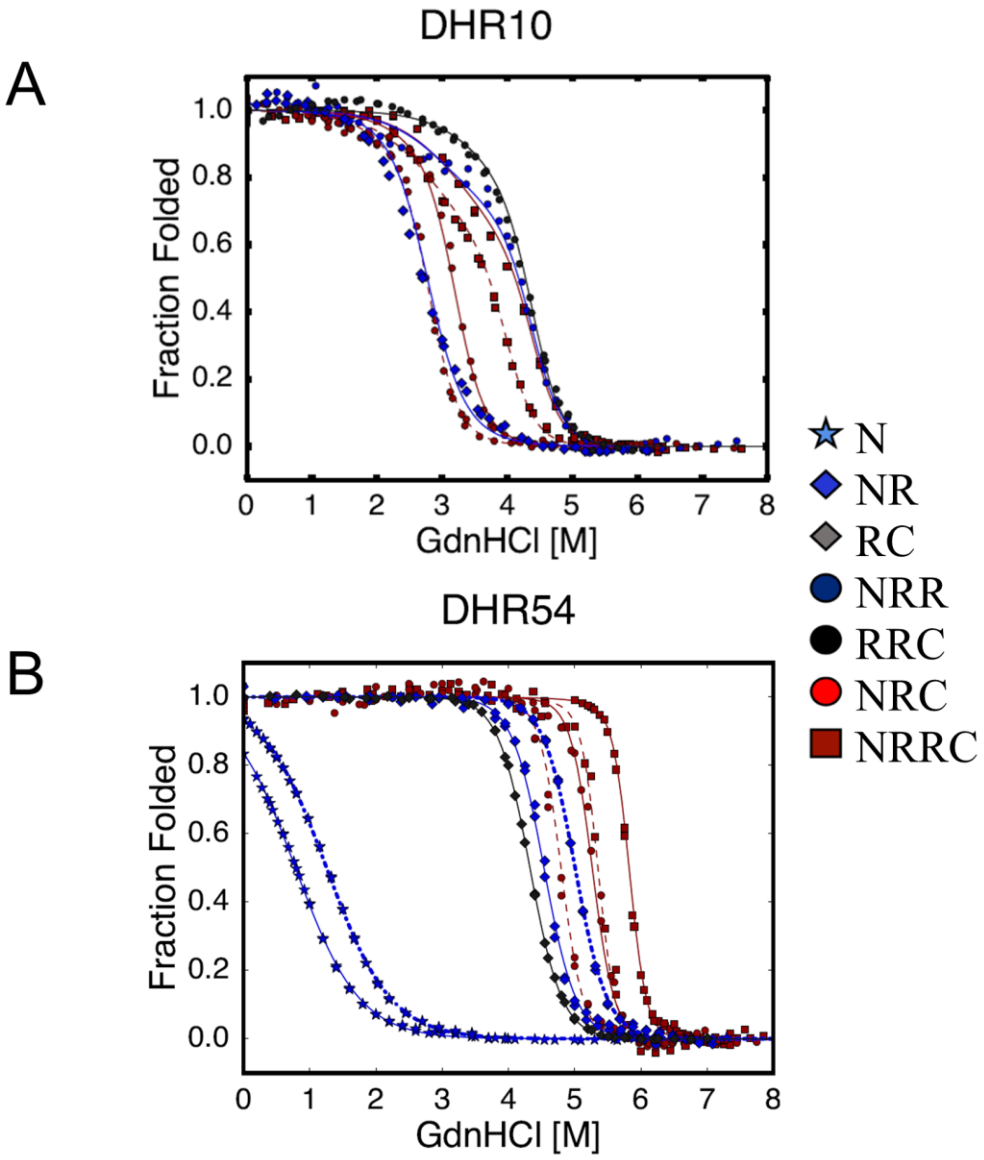


Figure 3.2 Unfolding transitions and nearest-neighbor Ising analysis of DHR proteins of different length and capping architecture. Guanidine-induced unfolding transitions were fitted with a nearest-neighbor Ising model (curves). N-capped constructs are shown in blue, C-capped constructs are shown in grey, and doubly-capped constructs are shown in red. Differences in glycerol concentrations are shown using different line styles: 0% glycerol, dash-dotted curves; 10% glycerol, solid curves; 20% glycerol, dashed curves). For all constructs, increasing the number of repeats increases stability (based un unfolding midpoints). Likewise, increasing glycerol concentration increases stability, although glycerol stabilizes DHR10.2 (A) and DHR54 (B) to a greater extent than DHR71 (C) and DHR79 (D). Conditions: 25 mM $NaPO_4$, 150 mM NaCl, 25°C. (continued on next page)

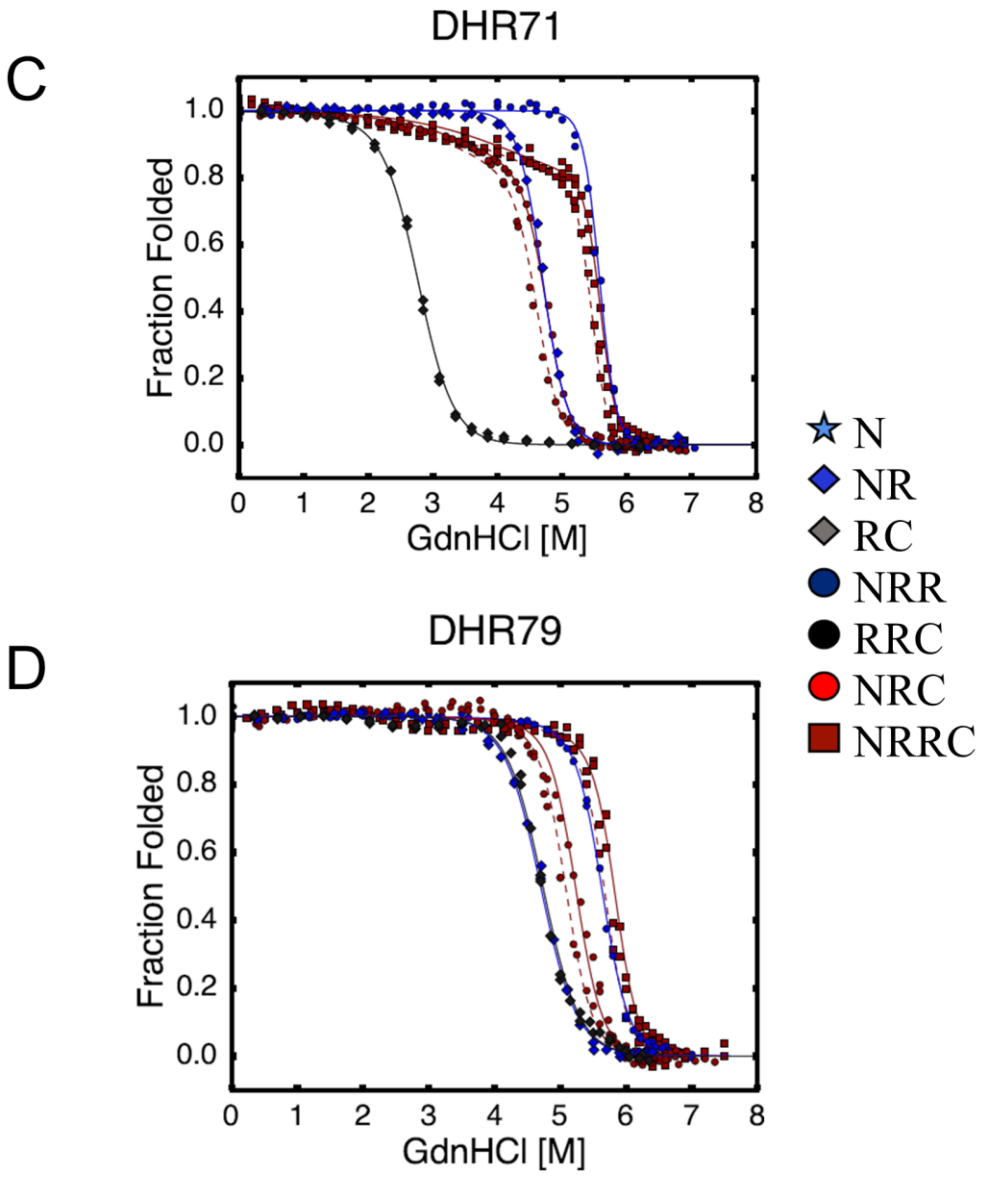


Figure 3.2 (continued) Unfolding transitions and nearest-neighbor Ising analysis of DHR proteins of different length and capping architecture. Guanidine-induced unfolding transitions were fitted with a nearest-neighbor Ising model (curves). N-capped constructs are shown in blue, C-capped constructs are shown in grey, and doubly-capped constructs are shown in red. Differences in glycerol concentrations are shown using different line styles: 0% glycerol, dash-dotted curves; 10% glycerol, solid curves; 20% glycerol, dashed curves). For all constructs, increasing the number of repeats increases stability (based un unfolding midpoints). Likewise, increasing glycerol concentration increases stability, although glycerol stabilizes DHR10.2 (A) and DHR54 (B) to a greater extent than DHR71 (C) and DHR79 (D). Conditions: 25 mM $NaPO_4$, 150 mM NaCl, 25°C.

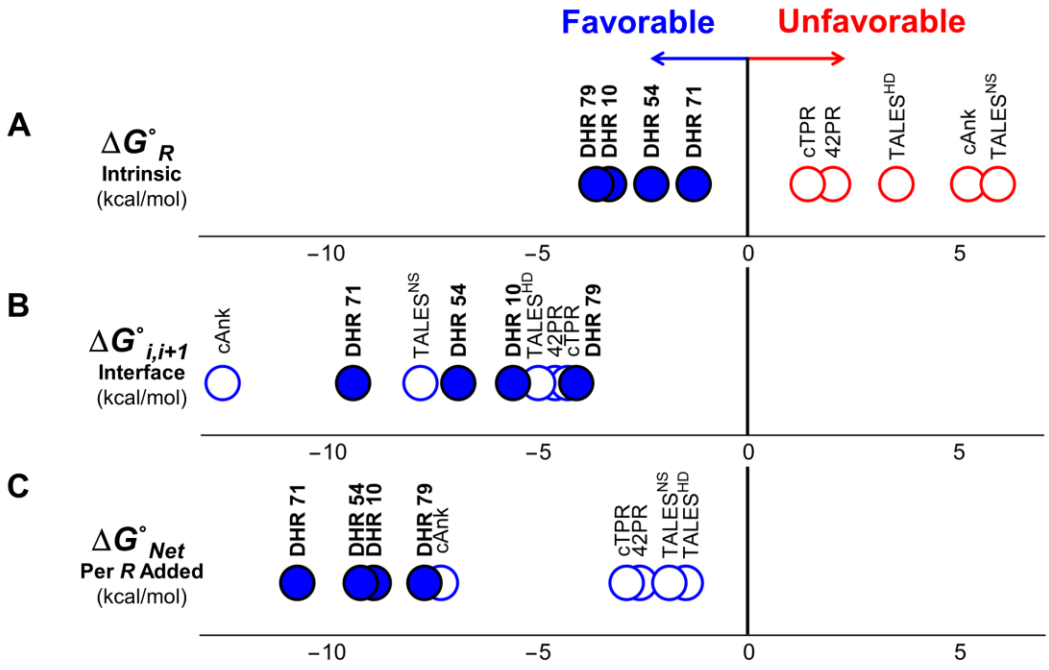


Figure 3.3 DHR repeats are intrinsically stable, unlike the repeats of naturally occurring repeat proteins. (A) Intrinsic and (B) interfacial coupling free energies determined by Ising analysis for designed helical repeat proteins (filled circles, this study) and natural repeat proteins (open circles, TALES^{NS} and TALES^{HD} (Geiger-Schuller and Barrick, 2016), 42PR (Marold et al., 2015), cANK (Aksel et al., 2011), cTPR (Marold et al., 2015)). Unfavorable (i.e., positive) free energy terms are in red, favorable (i.e., negative) folding free energies are in blue. Designed helical repeats are stabilized by both favorable intrinsic and interfacial coupling folding free energies, while natural repeat proteins are destabilized by unfavorable intrinsic folding free energies, which partly offset favorable interfacial interactions. (C) Free energy associated with adding a single repeat to a folded array (the sum of free intrinsic and interfacial free energies in panels A and B). Due to both their favorable intrinsic folding free energies, DHR proteins are more strongly stabilized by the addition of repeats than natural repeat proteins, and as a result, are extraordinarily stable.

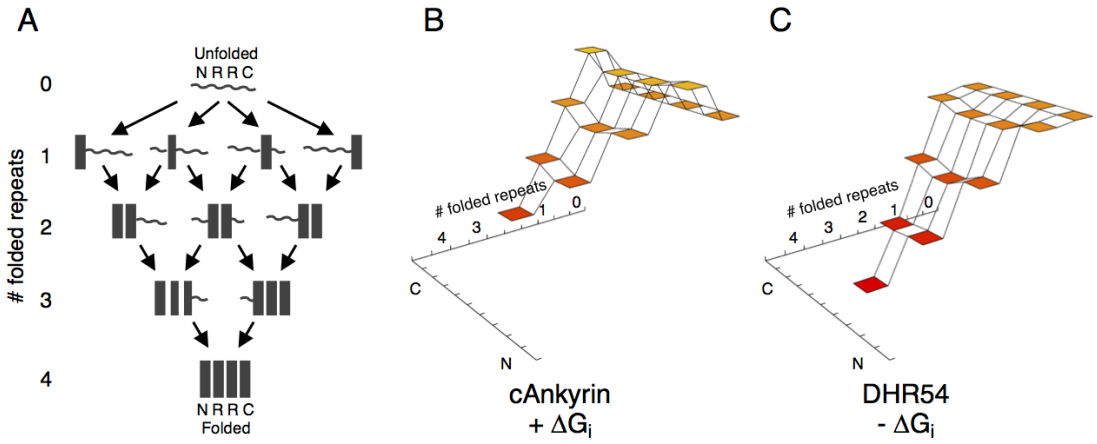


Figure 3.4 Stabilizing intrinsic energies create barrierless folding energy landscapes for DHR proteins in the absence of denaturant. (A) Repeat proteins with NR_2C repeat sequences can fold along many pathways. (B and C) Free energy landscapes from experimentally determined intrinsic and interfacial free energies. The vertical dimension (and shading) shows the free energies of partly folded states along the folding pathway shown in (A). (B) Consensus ankyrin repeat proteins, which are based on the naturally occurring ankyrin repeat family, have destabilizing intrinsic energies, and as a result, folding the first repeat results in a barrier to folding. (C) DHR54 proteins have stabilizing intrinsic folding energies, folding the first repeat is energetically favorable, and addition of subsequent repeats are strongly down-hill. Landscapes were generated with the "Graphics" command of Mathematica.

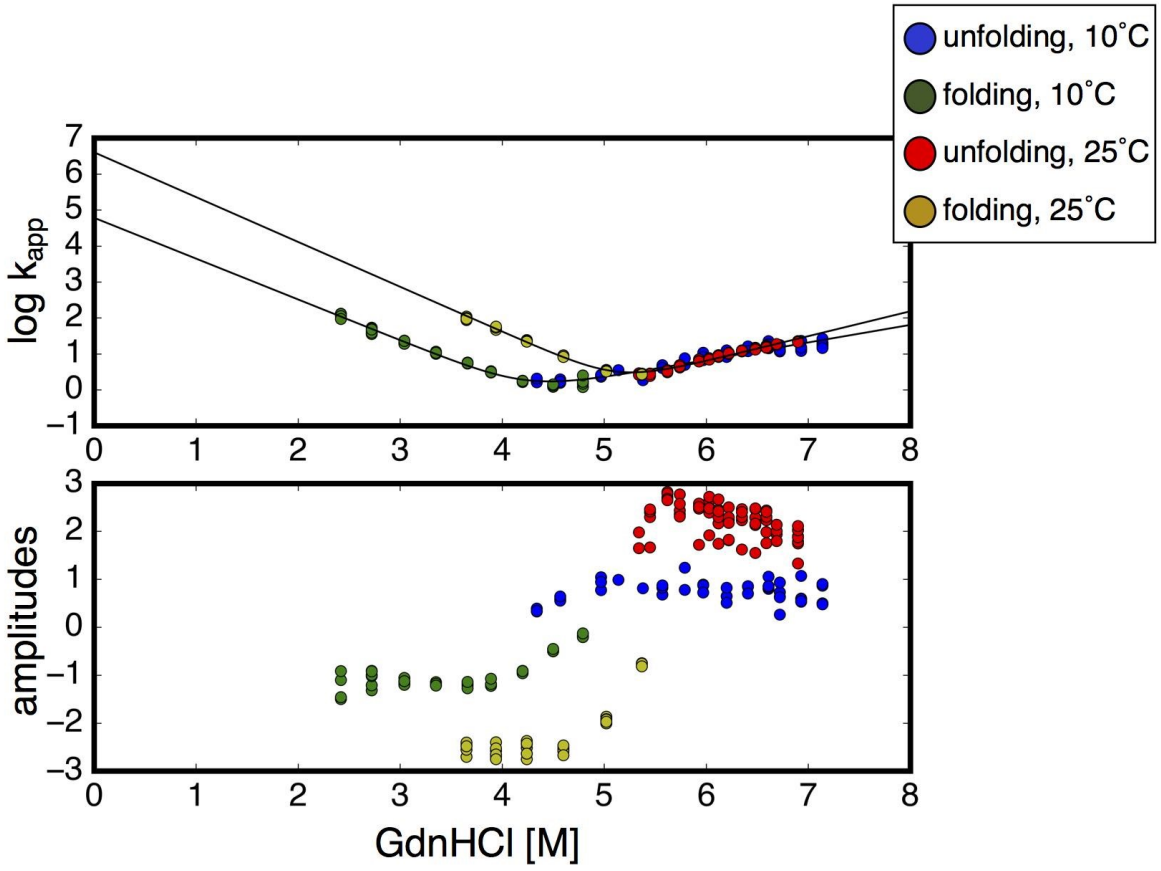


Figure 3.5 Folding kinetics of DHR54 NRC. Stopped-flow fluorescence folding kinetics measured at 25°C (containing 10% glycerol) and 10˚C (no glycerol). The log of apparent rate constants fitted from 3 traces shown in the top panel (circles). Amplitudes (ΔY in Equation 3.4) shown in circles below. Data are colored by temperature and folding direction (see legend). In addition to glycerol, samples contained 25 mM NaPO4, 150 mM NaCl.

Table 3.1. Thermodynamic parameters obtained from Ising fits.

	ΔG_N	ΔG_R	ΔG_C	$\Delta G_{i,i+1}$
DHR10.2	1.46 [1.26, 1.67]	-2.51 [-2.90, -2.15}	0.63 [0.32, 1.00]	-4.80 [-5.10, -4.53]
DHR54	-0.45 [-0.58, -0.32]	-2.04 [-2.17, -1.92]	-0.84 [-0.94, -0.74]	-6.76 [-6.98, -6.54]
DHR71	-3.01 [-3.27, -2.75]	-1.41 [-1.61, -1.23]	3.06 [2.87, 3.29]	-9.93 [-10.50, -9.43]
DHR79	-1.84 [-2.06, -1.64]	-3.48 [-3.83, -3.22]	-1.81 [-2.08, -1.61]	-4.83 [-5.14, -4.55]

	$m_{GdnHCl,\ i}$	$m_{Glycerol,\ i}$	$m_{GdnHCl,\ C}$	$\Delta G_{N,i+1}$
DHR10.2	-1.23 [-1.33, -1.14]	0.36 [0.33, 0.40]	N/A	N/A
DHR54	-1.24 [-1.28, -1.21]	0.41 [0.39, 0.43]	N/A	-7.72 [-7.95, -7.49]
DHR71	-1.57 [-1.66, -1.49]	0.17 [0.15, 0.20]	-0.71 [-0.79, -0.64]	N/A
DHR79	-1.12 [-1.18, -1.06]	0.15 [0.12, 0.18]	N/A	N/A

Free energies have units of kcal/mol. m_{GdnHCl} and $m_{Glycerol}$ have units of kcal/mol/[M GdnHCl] and kcal/mol/[M Glycerol].

Table 3.2. Kinetic parameters obtained from DHR54 NRC stopped-flow analysis.

Temp.	$k_{f,\ H2O}$ [a]	$k_{u,\ H2O}$ [a]	m_f [b]	m_u [b]	ΔG [c]
10˚C	$6.1 \times 10^{4} \pm 1.8 \times 10^{4}$	$7.9 \times 10^{-3} \pm 2.1 \times 10^{-3}$	-1.13 ± 0.04	0.49 ± 0.02	8.92
25˚C	$4.0 \times 10^{6} \pm 1.0 \times 10^{6}$	$4.5 \times 10^{-4} \pm 1.1 \times 10^{-4}$	-1.25 ± 0.03	0.69 ± 0.02	13.58

[a] Rate constants have units 1/sec. Uncertainties are estimated from the covariance matrix (lmfit). [b] m_f and m_u have units of 1/[M GdnHCl]. [c] Free energies have units of kcal/mol and are calculated using Equation 3.6.

3.4 Discussion

By measuring the length-, capping-, and glycerol-dependence on stability of four DHRs families, we obtained intrinsic folding free energies and interfacial interaction free energies. Unlike previously studied consensus and natural proteins, we find that all DHRs have stabilizing intrinsic folding free energies. These results suggest the Rosetta algorithm and the methods used by Baker and coworkers to design the DHR proteins are particularly good at optimizing local stability. The very high stability of these DHR constructs results from this feature, performing no better on average (and no worse) than nature in terms of long-range coupling. Because repeats are intrinsically stable, it should be possible to mix different DHR types in one array and produce folded proteins.

3.4.1 Rosetta algorithms design stable proteins through favorable local interactions

Proteins designed from the Rosetta suite of algorithms are often hyper-stable (Brunette et al., 2015). How do these algorithms achieve such stability? There are three possibilities: First- Rosetta makes very stable units of structure (ex: very good helices); Second- Rosetta makes very favorable interfaces that

tightly couple adjacent units of structure; Third- a combination of the previous possibilities. In this work, we dissect the stability of many DHRs into intrinsic energies and coupling energies to address the origins of this hyper-stability.

3.4.2 Favorable local interactions of DHRs reduce folding barriers

Figure 3.4A depicts possible folding pathways for a four repeat protein (NR_2C). Ignoring lower probability configurations where one unfolded repeat is flanked by two folded repeats, there are ten configurations along these pathways. Using published Ising parameters from a consensus Ankyrin repeat protein (cAnk), the energy of each microstate is calculated in Figure 3.4B (Aksel et al., 2011). Because the intrinsic energy of these naturally-derived repeats is unfavorable, all microstates with one folded repeat are unfavorable. While these single-repeat-folded microstates are high energy, they must be accessed on the way to stable microstates with two or more consecutive, coupled, folded repeats.

Using fitted Ising parameters from Table 3.1, we calculated the same type of landscape for DHR54 (Figure 3.4C). Because the intrinsic folding energy of these designed repeats is favorable, all partly-folded configurations are stable. Not only is the landscape for this designed protein smooth, it is also very steep, indicating a strongly downhill folding scenario under native conditions. Given

these landscape features, DHR54 should fold much faster than cAnk. Indeed, we find evidence for fast folding (Figure 3.5, Table 3.2), with extrapolated folding rate constant in water is 4.3×10^{6} sec^{-1}. The extrapolated unfolding rate constant in water is 2.8×10^{-4} sec^{-1}. Although this is a long extrapolation, we note that the simple two-state folding mechanism assumed by Equation 3.5 gives a predicted equilibrium stability and denaturant dependence that matches the equilibrium DHR54 values reasonably well (Table 3.2). The extrapolated rate constant for DHR54 folding is three orders of magnitude faster than the published folding rate constant of cAnk (Aksel and Barrick, 2014), consistent with folding energy landscapes shown in Figure 3.4. The fast rate of folding and simple mechanism are at odds with predictions from energy landscape theory (Watters et al., 2007), which postulate that simple efficient folding requires extensive evolutionary selection to minimize frustration.

3.5 Materials and Methods

3.5.1 Cloning, expression, and purification

Genes containing DHR repeat constructs were purchased as GeneStrings from GeneArt and cloned with C-terminal His_6 tags via Gibson Assembly. DHR constructs were grown in BL21(T1R) cells at 37°C to an OD of 0.6-0.8, induced with 0.2 mM IPTG, and expressed overnight at 17°C. Following cell pelleting, resuspension, and lysis, proteins were purified by affinity chromatography on a Ni-NTA column. Protein was eluted using 250 mM imidazole and dialyzed into 150 mM NaCl, 0-20% glycerol, and 25 mM Na_3PO_4 pH 7.0.

3.5.2 Circular Dichroism (CD) spectroscopy

Circular Dichroism measurements were collected using an AVIV model 400 CD Spectrometer (Aviv Biomedical, Lakewood, NJ, USA). Far-UV CD scans were collected at 25°C using a 0.1 cm pathlength quartz cuvette, with protein concentrations of 15-30 μM. Buffer scans were recorded and were subtracted from the raw CD data. CD-monitored guanidine unfolding transitions at 222

nm were generated with an automated titrator using 1.5-3 μM protein and a 1 cm pathlength quartz cuvette.

3.5.3 Ising analysis of DHRs

To determine the intrinsic and interfacial free energies for folding of DHR arrays, and to analyze energies of partly folded states, we used a one-dimensional Ising formalism (Aksel and Barrick, 2009; Poland and Scheraga, 1970). In this model, intrinsic folding and interfacial interaction between nearest neighbors are represented using equilibrium constants κ and τ, respectively, where

$$\kappa_N = e^{-\left(\Delta G_N - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (3.1.1)$$

$$\kappa_R = e^{-\left(\Delta G_R - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (3.1.2)$$

$$\kappa_C = e^{-\left(\Delta G_C - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (3.1.3)$$

$$\tau = e^{-\left(\Delta G_{R,\,i-1}\right)/RT} \quad (3.1.4)$$

For all DHRs, the intrinsic folding free energies of N (solubilizing N-terminal cap), R (consensus repeat), and C (solubilizing C-terminal cap) are each considered to be independent adjustable parameters. DHR10.2, DHR71, and DHR79 are well described by a simple model where the interfacial interactions of the N:R , R:R, and R:C, pairs are identical. DHR54 unfolding

transitions are better fitted by a model where the interfacial interactions of the R:R and R:C interface are identical, whereas the N:R pair is considered is independent. Glycerol and denaturant dependences are built into the intrinsic (but not the interfacial) terms. DHR71 unfolding transitions are better fitted by a model that includes a separate denaturant dependence for the C-terminal cap ($m_{GdnHCl,\,C}$, Table 3.1).

Using these equilibrium constants, a partition function q for an n-repeat construct can be constructed by multiplying two-by-two transfer matrices:

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau & 1 \\ \kappa_N & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau & 1 \\ \kappa_R & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C \tau & 1 \\ \kappa_C & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{3.2}$$

This representation correlates the each repeat to its neighbor through the separate rows of each matrix. The fraction folded (f_{folded}) can be expressed using differentiation:

$$f_{folded} = \frac{1}{nq}\left(\kappa_N \frac{\partial q}{\partial \kappa_N} + \kappa_R \frac{\partial q}{\partial \kappa_R} + \kappa_C \frac{\partial q}{\partial \kappa_C} \right) \tag{3.3}$$

Ising parameters were determined by globally fitting Eq. 3 to guanidine-induced unfolding transitions collected at 0, 10, and 20% glycerol. Fitting was performed using the nonlinear least squares algorithm of the lmfit package (Newville et al., 2014) using an in-house python program (written by J. Marold (Marold et al., 2015) and adapted to include glycerol dependence by K.Geiger-Schuller) Confidence intervals (95%) were determined by performing 2000 bootstrap iterations.

3.5.4 Stopped-flow fluorescence spectroscopy

Folding kinetics was measured on Applied Photophysics SX.18MV-R stopped-flow fluorometer (Leatherhead, UK) using final protein concentrations of 3-5 μM. Total tryphtophan fluorescence was collected using a 320 nm cut-off filter, following excitation at 280 nm. Single traces were fitted to obtain folding and unfolding rate constants using the following equation:

$$Y_{obs} = Y_{\infty} + \Delta Y e^{-k_{app}t} \tag{3.4}$$

Folding and unfolding rate constants at different guanidine-HCl concentrations were fitted globally with the following equation (Tripp et al., 2017):

$$\log k_{app} = \log(k_{f,H_2O} 10^{m_f [GdnHCl]} + k_{u,H_2O} 10^{m_u [GdnHCl]}) \tag{3.5}$$

where k_{f,H_2O} and k_{u,H_2O} are the folding and unfolding rate constants in the absence of guanidine (respectively). Unfolding free energy estimated using the following equation:

$$\Delta G = RT ln(\frac{k_{f\,\mathrm{H20}}}{k_{u\,\mathrm{H20}}}) \tag{3.6}$$

Best-fit parameters and uncertainties estimated from the covariance matrix shown in Table 3.2.

3.6 References

Aksel, T., and Barrick, D. (2009). Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. Methods Enzymol. 455, 95–125.

Aksel, T., and Barrick, D. (2014). Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. Biophys. J. 107, 220–232.

Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993 19, 349–360.

Brunette, T.J., Parmeggiani, F., Huang, P.-S., Bhabha, G., Ekiert, D.C., Tsutakawa, S.E., Hura, G.L., Tainer, J.A., and Baker, D. (2015). Exploring the repeat protein universe through computational protein design. Nature 528, 580–584.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. Genetics 186, 757–761.

Cortajarena, A.L., Kajander, T., Pan, W., Cocco, M.J., and Regan, L. (2004). Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. Protein Eng. Des. Sel. PEDS 17, 399–409.

Cunha, E.S., Hatem, C.L., and Barrick, D. (2016). Synergistic enhancement of cellulase pairs linked by consensus ankyrin repeats: Determination of the roles of spacing, orientation, and enzyme identity. Proteins 84, 1043–1054.

Geiger-Schuller, K., and Barrick, D. (2016). Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. Biophys. J. 111, 2395–2403.

Kajander, T., Cortajarena, A.L., Main, E.R.G., Mochrie, S.G.J., and Regan, L. (2005). A new folding paradigm for repeat proteins. J. Am. Chem. Soc. 127, 10188–10190.

Kloss, E., Courtemanche, N., and Barrick, D. (2008). Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. Arch. Biochem. Biophys. 469, 83–99.

Li, T., Huang, S., Jiang, W.Z., Wright, D., Spalding, M.H., Weeks, D.P., and Yang, B. (2011). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res. 39, 359–372.

Marold, J.D., Kavran, J.M., Bowman, G.D., and Barrick, D. (2015). A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Struct. Lond. Engl. 1993.

Newville, M., Stensitzki, T., Allen, D.B., and Ingargiola, A. (2014). LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python¶ (Zenodo).

Poland, D., and Scheraga, H.A. (1970). Theory of helix-coil transitions in biopolymers: statistical mechanical theory of order-disorder transitions in biological macromolecules (Academic Press).

Tripp, K.W., Sternke, M., Majumdar, A., and Barrick, D. (2017). Creating a Homeodomain with High Stability and DNA Binding Affinity by Sequence Averaging. J. Am. Chem. Soc. 139, 5051–5060.

Watters, A.L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., and Baker, D. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell 128, 613–624.

Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K., and Plückthun, A. (2008). Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. 376, 241–257.

Chapter 4

Probing the local stability dependence of a processive protease

4.1 Introduction

Rates of proteolysis can be tightly linked to the thermodynamic stability of target protein substrates. As protein stability increases, susceptibility to degradation decreases. The bacterial thermostable metalloprotease thermolysin has been successfully used to probe regions of protein instability *in vitro* (Park and Marqusee, 2005). This method of pulse-proteolysis exploits the ability of housekeeping proteases like thermolysin to peck non-specifically at hydrophobic residues in protein regions whose folding equilibrium is biased toward local unfolding.

However, the main mechanism of protein quality control in living cells occurs through closely linked protein chaperones and proteasome components. In eukaryotes and prokaryotes, ATPases Associated with diverse cellular Activities, known as AAA^+ enzymes, contain the majority of these protease and chaperone complexes such as ClpXP, ClpAP, ClpCP, HslUV, Lon, FtsH, PAN/20S, and components of the 26S proteasome. Proteases from this class degrade target protein substrates through processive degradation mechanisms that are tightly coupled to ATP hydrolysis. The hydrolysis of ATP provides the energy necessary to disrupt higher order protein structures and shuttle unfolded substrates through the enzyme interior (Gur et al., 2012). Internal catalytic peptidase residues hydrolyze peptide bonds processively to fully degrade the target substrate.

Proteolysis by AAA+ proteases has been shown to target both misfolded proteins and specific target substrates. The latter targets are recognized by sequence-specific motifs, known as degrons. For example, β-galactosidase degradation is initiated by recognition by the Lon protease of exposed unfolded N-terminal sequences. These residues have been successfully used to design the 20 amino acid β20-degron sequence to target foreign substrates for degradation studies.

While global thermodynamic stability of target substrates has been shown to affect rates of proteolysis by processive proteases (Kenniston et

al., 2003), the extent to which local stability impacts proteolysis rates is not known. Protease action is in constant battle between unfolding and refolding of the target substrate (Martin et al., 2007). For cells, it is energetically cheaper to allow misfolded proteins to refold than actively degrade them. However, once target degradation signals are bound and processed through the protease core, the protease ATPase units continue to drive the folding equilibrium to the unfolded state. Thus it follows that perturbations in this folding equilibrium by local changes in stability would result in comparable changes in degradation rate. How do local variations in stability near or far from degradation initiation affect rates of proteolysis? At what distance from the degron tag do proteases sense changes in substrate stability? To examine the interplay between changes in local stability and protein degradation by processive proteases, we have designed a degron-tagged consensus ankyrin repeat protein substrate for *in vitro* cellular protease assay with *E. coli* Lon protease, based on previous studies examining the mechanism and action of Lon protease (Gur and Sauer, 2008). Using point mutations, the local stability of the ankyrin repeat substrate can be modulated, allowing us to test the effects of local stability variation both near to and far from the degron tag.

Lon protease is a bacterial protease that contains both the ATPase and peptidase domains on the same translated peptide. At cellular protein concentrations, Lon protease monomers form a homohexameric

ring complex. This quaternary organization is necessary for functional degradation of substrates, and higher and lower order protease complexes cannot efficiently act to the same extent (Vieux et al., 2013). This simplifies biochemical reagent preparations, compared to similar ClpXP complexes that rely on the oligomerization of a homohexameric ClpX ATPase and a tetradecameric ClpP endopeptidase (Baker and Sauer, 2012). The use of a fused ATPase and peptidase protease allows us to infer coupled degradation with protein unfolding.

Consensus ankyrin repeat proteins are linear arrays of repeating helix-loop-helix motifs that are designed from the multiple sequence alignment of the ankyrin domain. The resulting consensus protein sequence has been shown to be more stable than naturally occurring ankyrin repeats (Tripp and Barrick, 2007), and studies that vary repeat number have resolved the intrinsic and interfacial contributions to the thermodynamic fold stability of the protein (Aksel et al., 2011). Furthermore, we have made a number of amino acid substitutions away from the consensus sequence and resolved the effect these substitutions have on intrinsic and interfacial stabilities (Chapter 2).

These consensus ankyrin repeat arrays are a model system for degradation studies of processive proteases for three reasons. First, the linear architecture of consensus ankyrin repeat arrays limits the interactions to adjacent repeats. In globular proteins, interactions can

(and often do, see (Bai et al., 1995)) occur via sequence-distant contacts. The linear architecture of ankyrin repeat arrays allows us to make local perturbations to stability at one location either proximal or distal to the fused terminal degron tag. Second, the relatively high stability of consensus ankyrin repeats provides for a robust background to withstand destabilizing mutations. Third, the resolution of intrinsic and interfacial folding contributions to stability allows us to better understand the stability environment down to the level of the individual repeat and interface.

4.2 Results

4.2.1 Developing an *in vitro* assay to study processive proteolysis

Here we seek to develop an *in vitro* assay to monitor degradation of target substrates by *Escherichia coli* Lon protease. While recombinant *E. coli* Lon protease can be expressed to high-levels under normal laboratory conditions, we experienced a number of problems purifying satisfactory amounts of protein. This is most likely due to either the large size of the Lon monomer (87.4 kDa) or the fact that Lon monomers have been shown to oligomerize into different quaternary structures (dimers, trimers, hexamers, and dodecamers) depending on concentration (Vieux et al., 2013). To separate sufficient amounts of pure Lon enzyme, we used a larger His_{12} purification tag that can bind nickel with greater affinity.

Although the His_{12} tag did improve the purity of the protein that bound and eluted from the nickel column, we continued to lose a large amount of protease in the flow-through and column wash. Suspecting

that oligomers of Lon may occlude the His-tag, we investigated the effects of different solvent conditions in the purification strategy to increase protein yield. While high concentrations (8M) of urea increased protein yield, most of the purified enzyme aggregated upon removal of denaturant either while on the affinity column or during dialysis. Refolding a denatured 87.4 kDa protein and restoring the enzyme to a fully functional state did not seem possible. However, the addition of 1M NaCl under native conditions increased protein yield to sufficient amounts (data not shown).

To-date, it is known that the class of AAA+ proteases, of which Lon is a member, can recognize a series of peptide degron tags, β20, sul20, and peptides with carboxymethylated cysteines. These degron tags have been previously characterized for different substrates (Gur and Sauer, 2009). Degron tags are effective at both the N- and C-termini, and also at internal sites within the protein primary structure. Because we sought to examine the extent to which local disruptions in protein structure and stability affect global degradation in a distance-dependent manner, we utilized only terminal degron tags, and did not introduce degrons within the primary structure.

We genetically fused a 20-amino acid sequence (QLRSLNGEWRFAWFPAPEAV) to the N-terminus of three-repeat consensus ankyrin, NRC. This degron tag, known as β20, is derived

from the N-terminus of β-galactosidase, and is responsible for the selective degradation of its native substrate. The β20 degron tag reveals a characteristic spectrum of a disordered protein by circular dichroism spectroscopy (Gur and Sauer, 2008). Far-UV CD spectra of NRC consensus ankyrin repeats and β20-tagged NRC retain the characteristic α-helical shape of ankyrin repeat arrays (Figure 4.2A). The molar residue ellipticity of β20-tagged NRC is diminished slightly compared to NRC, likely a result of the addition of 20 unstructured residues. This confirms that the β20 tag does not affect secondary structure of folded ankyrin repeats.

We next investigated the selectivity and functionality of purified recombinant *E. coli* Lon protease for degron-tagged substrates. To match physiologically relevant Lon protease concentration levels, we perform all experiments at 3.2 μM concentration of protease monomer, or 533 nM Lon hexamer. This has been previously shown to be the physiological cellular concentration of Lon monomer and is sufficient to form functional protease hexamers (Vieux et al., 2013).

It has been shown that Lon protease activity requires ATP binding and hydrolysis to unfold native substrates (Kenniston et al., 2003). Here we confirm that purified recombinant Lon protease selectively degrades β20-tagged NRC consensus ankyrin (Figure 4.1). Indeed, β20-BRC is degraded by Lon over the course of one hour. Here we find that this

reaction is ATP-dependent. Because untagged NRC is stable over this time course, we conclude that the β20 degron sequence is required for degradation of the substrate.

To evaluate the kinetics of substrate degradation, we monitored the *in vitro* reaction by SDS-PAGE, quenching reaction time point aliquots with Coomassie stain buffer containing reducing agent and SDS (Figure 4.3A). It should be noted that we observe the relatively slow disappearance of Lon protease as a function of time. It is unclear if this disappearance is attributed to self-degradation or protein adsorption to the reaction vessel walls over time. However, the timescale of this rate of disappearance is slow compared to the rate of disappearance of the intact substrate band.

To quantify the disappearance of intact substrate by SDS-PAGE, we use densitometry. Quantification of band intensities is performed by dividing the intact substrate band intensity by the pyruvate kinase (PK) loading control intensity. This ratio is then normalized by the ratio and substrate concentration at time = 0 min (Figure 4.3B). To obtain the initial rate of degradation (V_0), we fit the normalized band intensity ratios to an exponential decay and take the derivative to that curve at time = 0 min (Figure 4.3C).

To resolve the underlying parameters that contribute to V_0, we run the proteolysis reaction at different substrate concentrations (Figure 4.4).

Previous work (Gur and Sauer, 2008, 2009; Gur et al., 2012) has shown Lon protease degradation kinetics follow a modified form of the Michaelis-Menten equation ($V = V_{max} * [S]^n/(K_M^n + [S]^n)$ where n is a Hill coefficient.

4.2.2 Guanidine hydrochloride induced unfolding transitions of degron-tagged consensus ankyrin repeat proteins

The addition of an unstructured β20-degron has been shown to not interfere with global protein structure of consensus ankyrin repeats (Figure 4.2). We next investigated the extent to which unstructured β20 residues affect global protein stability. By monitoring guanidine-HCl induced unfolding transitions by far-UV CD at 222 nm, we find that the addition of unstructured β20 degron residues to the N-terminus of NRC does not affect global protein stability (Figure 4.2B). Both sets of data are fitted with a two-state model revealing fit global ΔG° values of 5.75 ± 0.19 kcal*mol^{-1} and 5.69 ± 0.09 kcal*mol^{-1} for NRC and β20-NRC respectively. An unpaired t-test reveals a two-tailed P-value of 0.65, indicating the two free energies are not statistically different.

Because the β20 tag does not affect folding stability, degron-tagged construct stabilities can be inferred from fit values for guanidine-HCl induced unfolding transitions of the untagged substrates, both using two-state and Ising parameters, for NR2C and for T4V variants (see Chapter 2).

4.2.3 Degradation kinetics of degron-tagged consensus ankyrin repeat variants

To investigate the extent to which local and global protein stabilities affect rates of proteolysis, we degraded β20-degron tagged consensus ankyrin repeats using our *in vitro* Lon proteolysis assay (Figures 4.4 and 4.6). Degradation of three repeat NRC consensus ankyrin occurs at a faster rate than the more stable four repeat NRRC (Table 4.1), consistent with the hypothesis that proteins with higher global stability are degraded slower than those with decreased stability. We find that less stable substrates allow for faster maximal rates of degradation by Lon protease, but the substrate K_M remains unchanged. The extent to which this increase in protein stability decreases rates of degradation is still to be determined. To investigate this, longer degron-

tagged repeat constructs with much higher stabilities will need to be degraded.

Degradation of degron-tagged N*-R*-R-C and N-R-R*-C* T4V substituted repeats reveals that both constructs are hydrolyzed at equal rates (Figure 4.6, Table 4.1). Since these rates are greater than that of four-repeat NRRC consensus ankyrin but less than three-repeat NRC consensus ankyrin, we conclude that lower global protein stability is related to faster degradation kinetics. However, it is apparent that the destabilizing substitutions localized to the N- or C-termini of four-repeat consensus ankyrin are not sensed by Lon protease. The ability of Lon protease to discern local stability changes in a distant-dependent manner most likely occurs at distances longer than four repeats (~40 Angstroms end to end, though the positions of substitution in our two T4V constructs differ by half that distance). To investigate this further, longer consensus repeat arrays with varying degrees of locally destabilizing substitutions should be assayed.

4.3 Figures and Tables

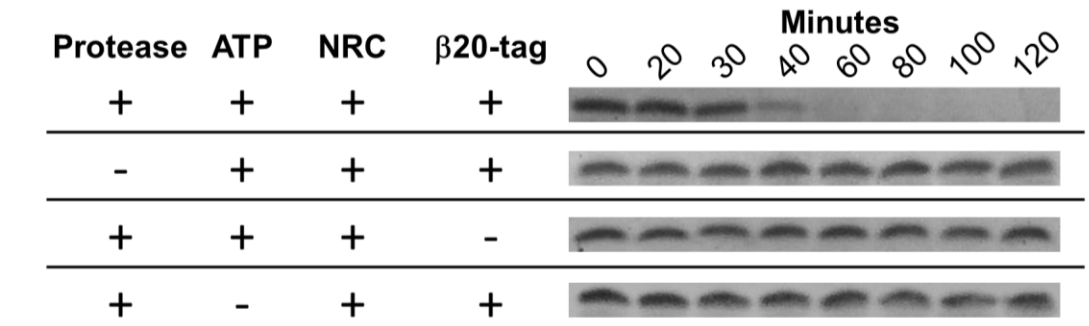


Figure 4.1 *E. coli* Lon protease selectively degrades degron-tagged substrates. Intact substrate is monitored by SDS-PAGE over time. Reaction controls shown top to bottom: absence of ATP, absence of Lon protease, absence of substrate-degron tag, and presence of all three (ATP, protease, and degron tag). Proteolytic degradation by Lon protease requires all three reaction conditions.

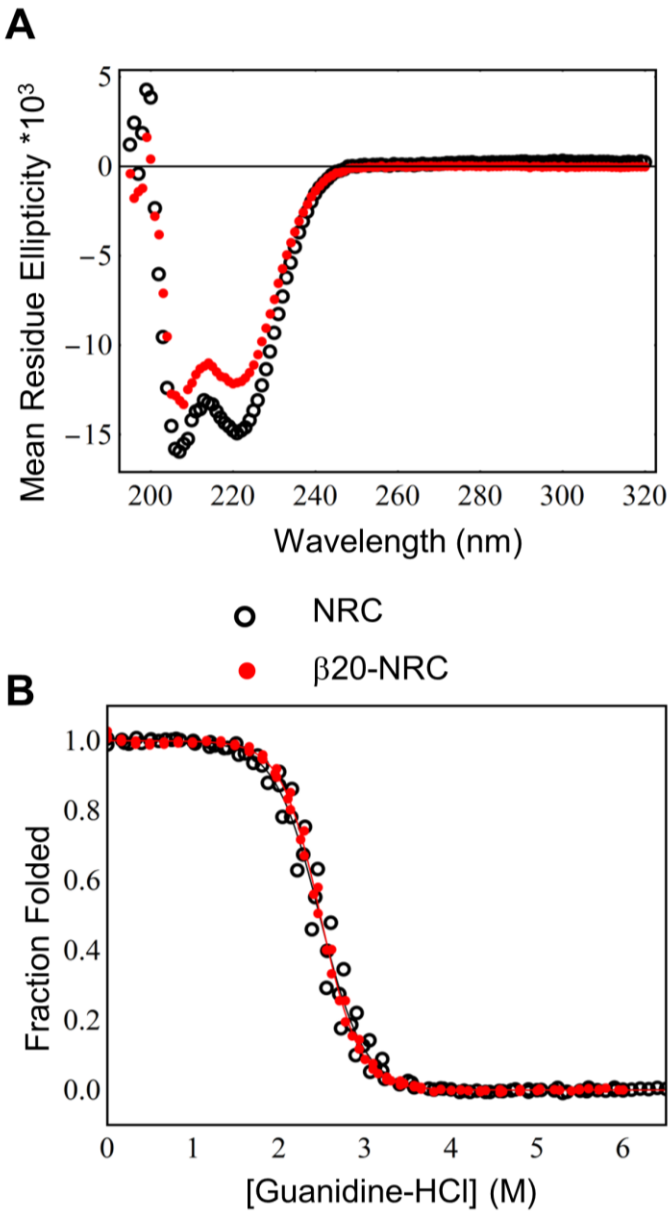


Figure 4.2 β20-degron tag does not interfere with substrate structure or stability. (A) Far-UV CD shows characteristic α-helical signal with minima at 208 nm and 222 nm. Differences in magnitude are likely to result from addition of the disordered degron-tag. (B) Guanidine-HCl induced unfolding transition of untagged and degron-tagged 3-repeat NRC consensus ankyrin monitored by CD spectroscopy at 222 nm. Solid curves are fits using a two-state model. Shown are duplicate data (circles) with a single average fit (line). Fitted baseline parameters were used to convert the data to fraction folded. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C.

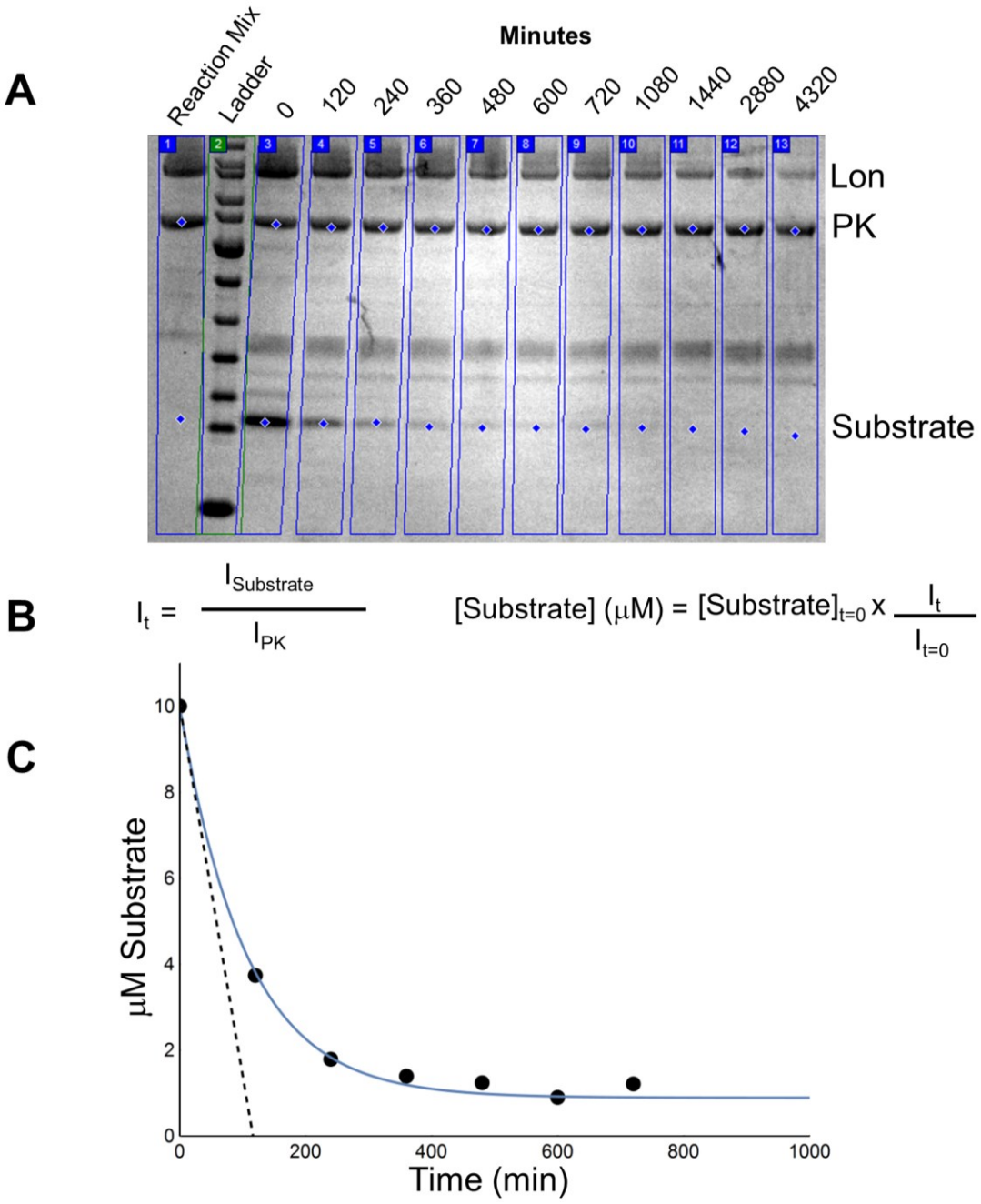


Figure 4.3 SDS-PAGE quantification of degron-tagged NRC by Lon protease. (A) Degradation of β20-tagged three-repeat NRC is monitored by SDS-PAGE. Top-most band is Lon-monomer, middle is pyruvate kinase, which is used as a loading control, bottom band is β20-NRC.. (B) Quantification of band intensities is performed by dividing the substrate band intensity by the PK loading control intensity. This ratio is normalized by the t_o ratio and the substrate concentration at t_o. (C) The initial rate is determined from the derivative at t_o (dashed line) of the single exponential fit (solid line) to band intensity data (black dots).

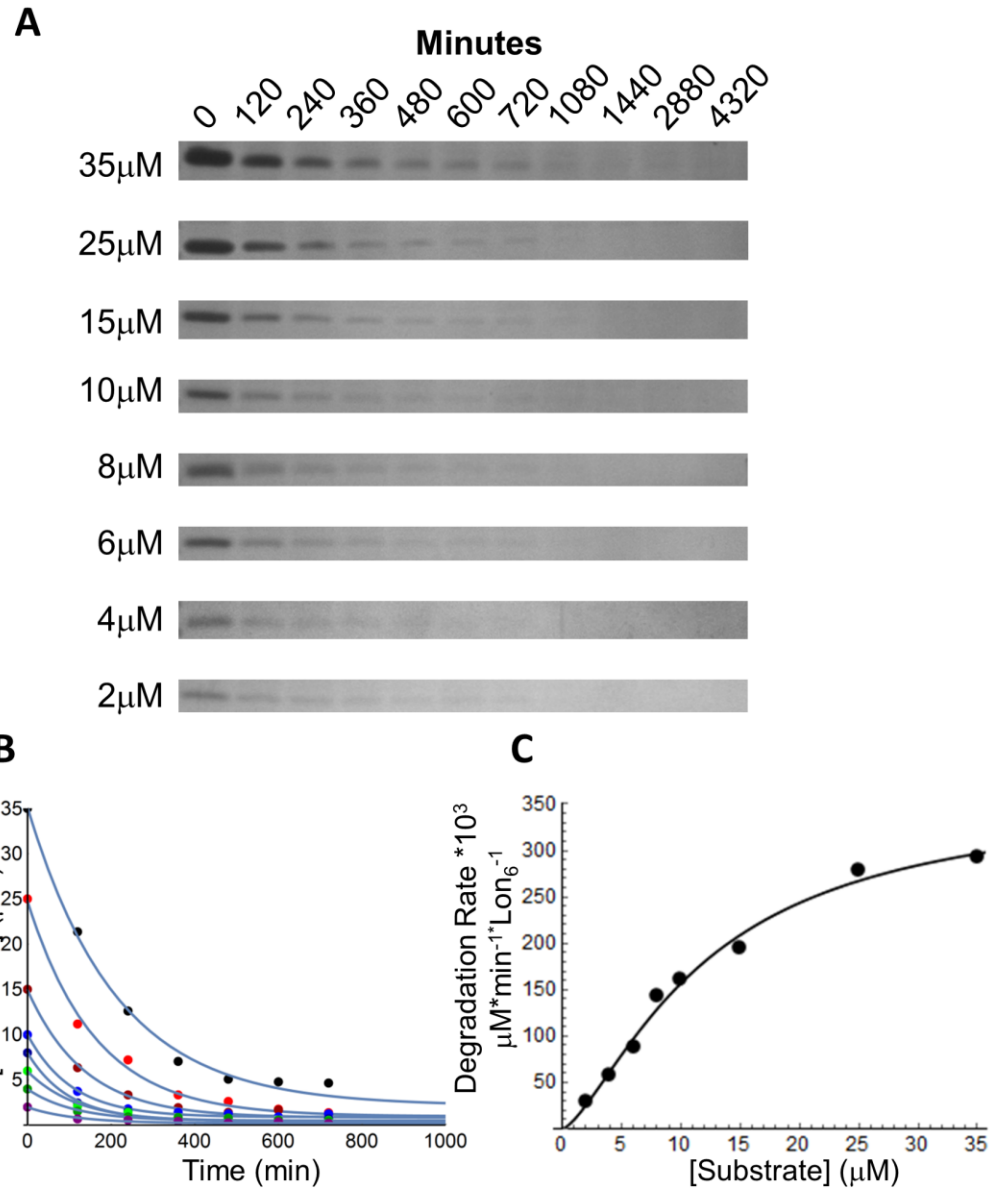


Figure 4.4 Degradation kinetics of Lon-digestion of β20-NRC. (A) β20-NRC is monitored over time by SDS-PAGE. Shown is the full-length NRC band; intensities are determined as in Figure 4.3, and plotted as a function of time in panel B at different starting substrate concentrations. Initial rates of degradation (determined from initial slopes in panel B) are shown in (C) fit with the Hill equation ($V_0 = V_{max} * [S]^n/(K_M^n + [S]^n)$).

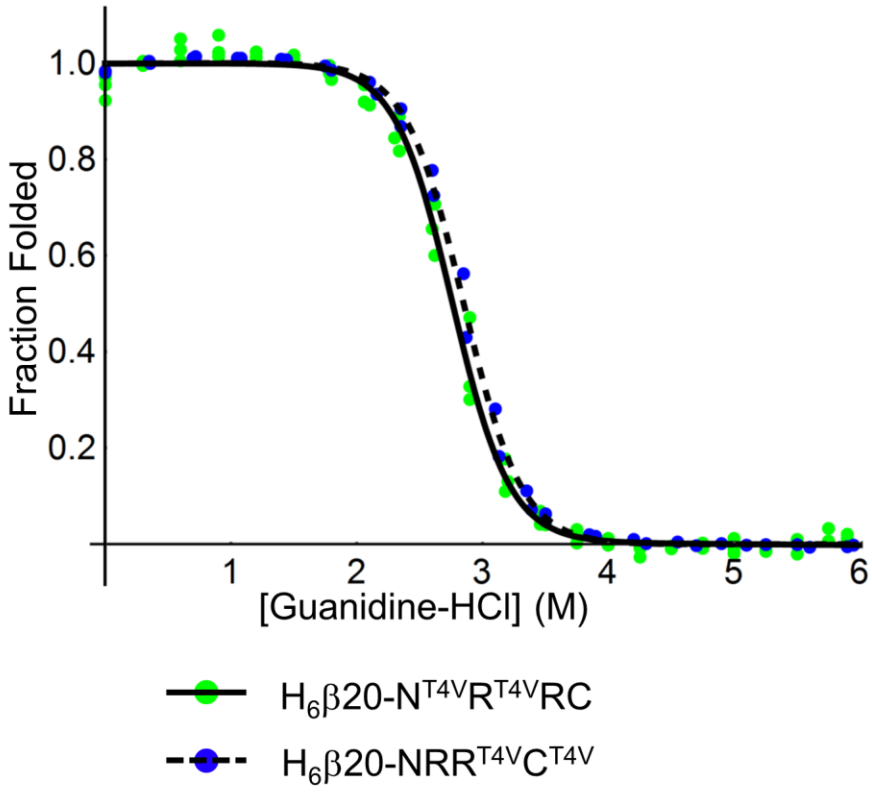


Figure 4.5 Equilibrium unfolding of T4V-substituded degron-tagged NRRC. T4V substitutions were made to the starred repeats of degron-tagged) N*R*RC and NRR*C* consensus ankyrin. Guanidine-HCl induced unfolding transition are monitored by CD spectroscopy at 222 nm. Shown are triplicate data (circles) and average fits for NRR*C* (dashed line) and N*R*RC (solid line). Data and fits are normalized after fitting by subtracting the fitted baselines. Conditions: 25 mM Tris-HCl pH 8.0, 150 mM NaCl, 20˚C.

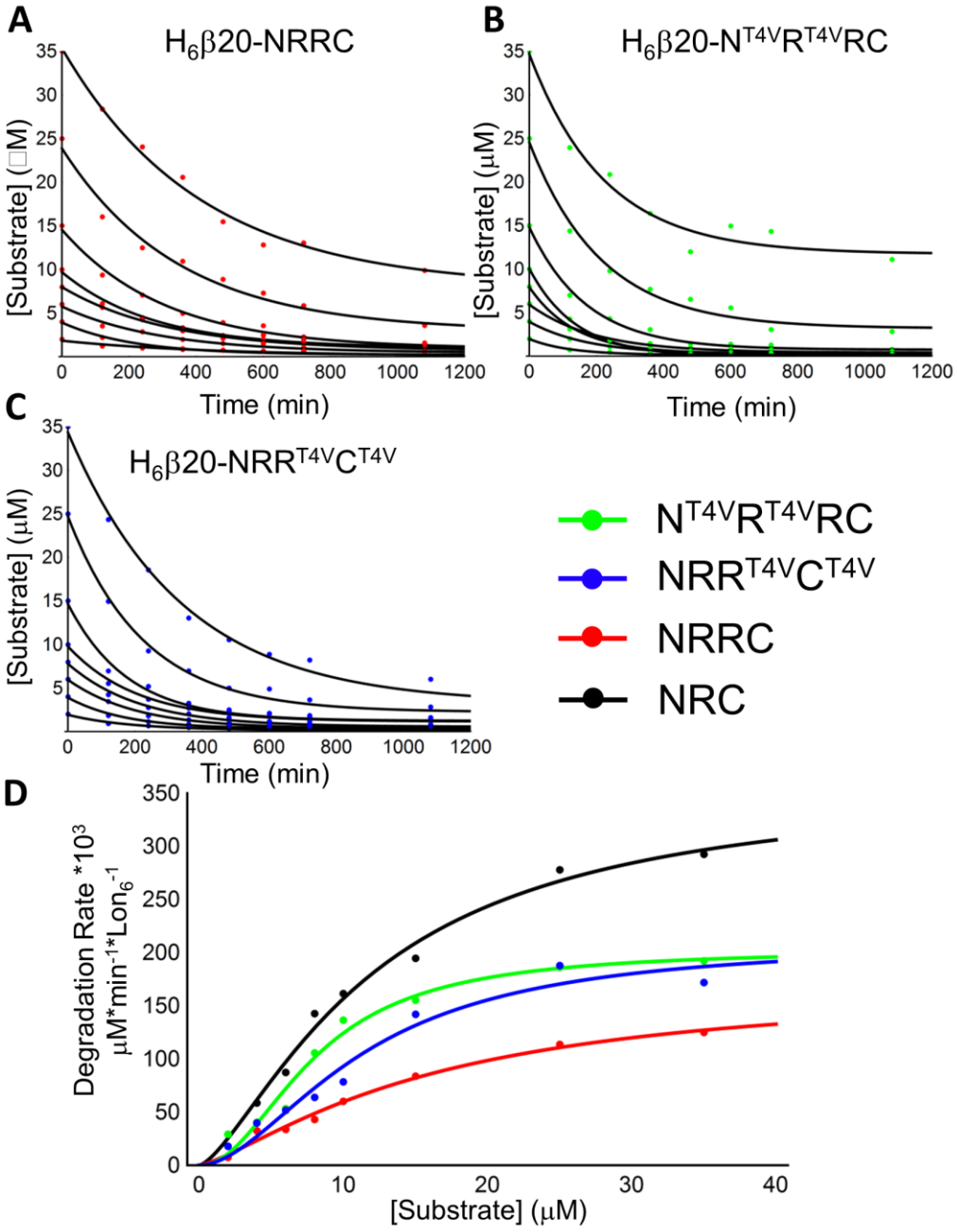


Figure 4.6 Kinetics of substrate degradation by Lon protease. Initial rate determinations of degradation of degron-tagged (A) NRRC in red, and T4V-substituted (B) N*R*RC in green and (C) NRR*C* in blue, as quantified by SDS-PAGE. (D) Initial rates are plotted and fit with the Hill equation ($V_0 = V_{max} * [S]^n/(K_M^n + [S]^n)$ and compared to NRC in black.

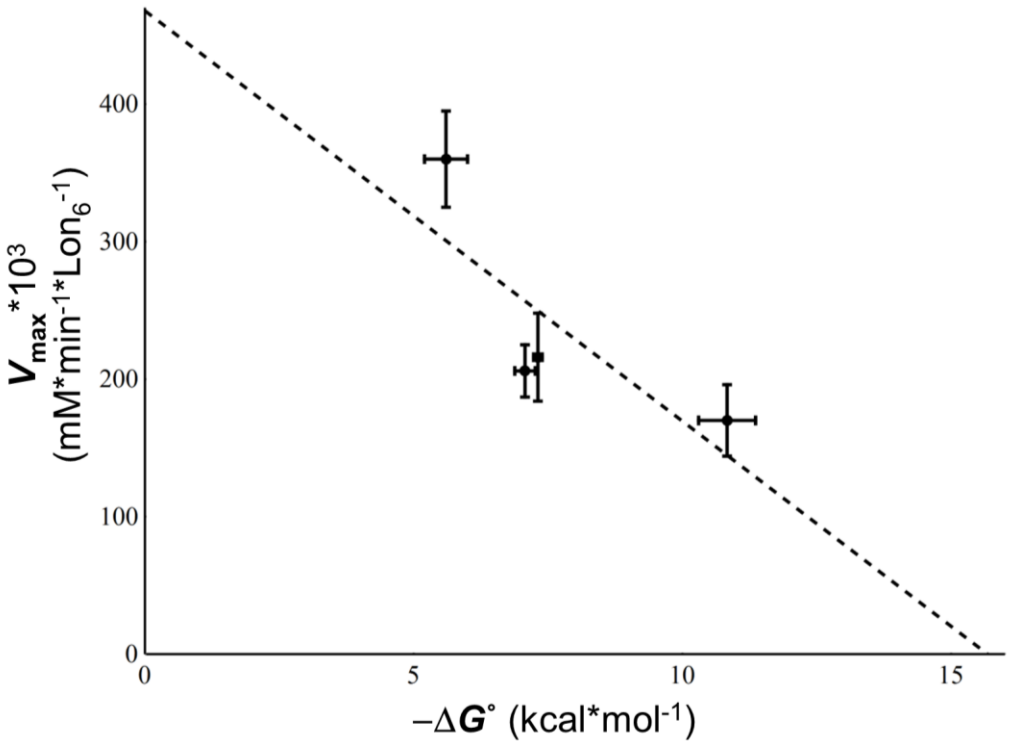


Figure 4.7 Rates of proteolytic degradation decrease with an increase in stability. Fit values for V_{max} are shown with respect to global stabilities. As substrate stability increases, degradation rates decrease, indicating that more work is required to degrade substrates with higher stabilities.

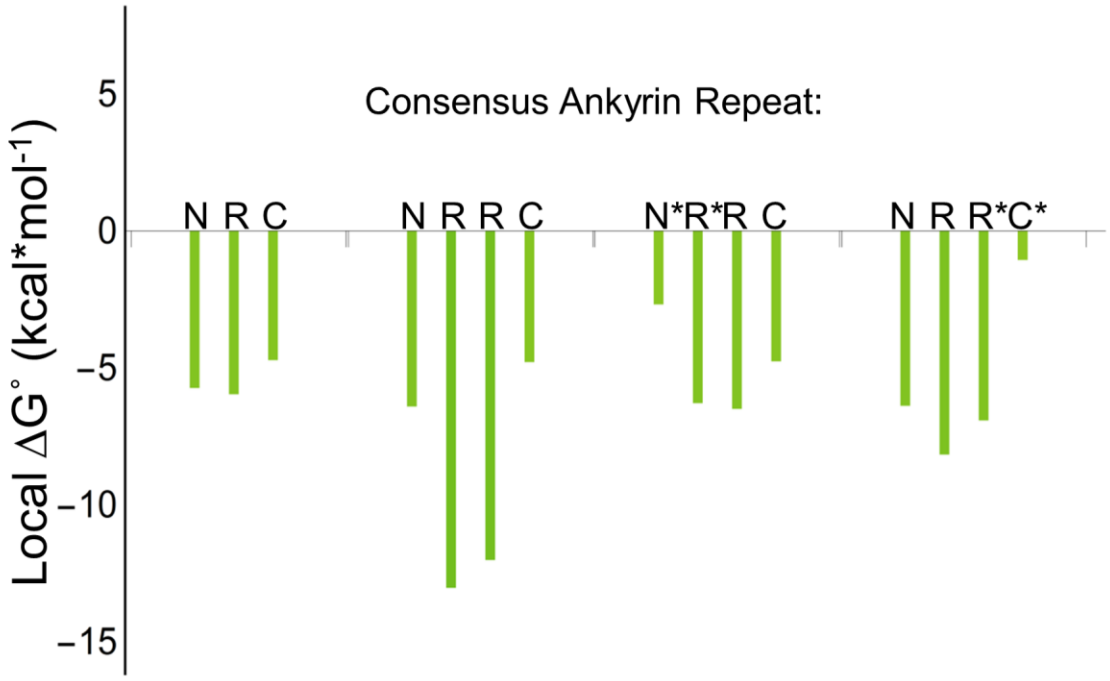


Figure 4.8 Distribution of local folding free energies for consensus ankyrin repeats of different lengths and repeat identity. At each repeat position, probabilities of being folded and unfolded were calculated from the nearest-neighbor partition function, using free energies from the Ising fit (Chapter 2). Free energies of local unfolding were calculated using the equation $\Delta G°_{local} = -RT\ln(p_U/p_F)$. The * indicates T4V substitution in the previous repeat.

Table 4.1 Ankyrin substrate stabilities and respective Lon proteolytic degradation kinetics parameters.

β20-tagged:	ΔG° * (kcal*mol^{-1})	V_{max}*10^{3} (μM*min^{-1} *Lon$_{6}^{-1}$)	K_M (μM)	***n***
NRC	-6.36 ± 0.12	360 ± 35	12.0 ± 2.3	1.4 ± 0.2
NRRC	-10.84 ± 0.54	170 ± 26	15.9 ± 4.6	1.3 ± 0.2
$N^{T4V}R^{T4V}RC$	-7.07 ± 0.19	206 ± 19	8.0 ± 1.2	2.0 ± 0.6
$NRR^{T4V}C^{T4V}$	-7.31 ± 0.08	216 ± 32	11.4 ± 3.3	1.8 ± 0.5

Kinetic parameters obtained from Hill equation fit to initial rate values. 95% confidence intervals obtained from bootstrap analysis (1000 iterations) assuming parameter uncertainties to be distributed normally.

*Free energy values determined from two-state fit of guanidine-HCl induced unfolding transitions monitored by CD spectroscopy at 222 nm.

4.4 Discussion

The motivation for this study is to better understand the relationship between protein stability and protein degradation in the cellular context. Cellular proteases are integral players in cell maintenance mechanisms and are responsible for protein turnover and quality control.

By developing an *in vitro* protease assay using *E. coli* Lon protease, we have begun to tease apart how thermodynamic stability correlates to susceptibility to proteolysis and protein turnover. We degraded three and four repeat consensus ankyrin protein arrays, and observed that with the stability increase conferred from the addition of a single repeat, the rate of degradation by Lon protease drastically decreases. By making destabilizing T4V substitutions of NRRC consensus ankyrin, we observe faster rates of protein degradation.

By comparing global ΔG° values obtained by guanidine-HCl induced unfolding transitions with relative rates of proteolysis, we see an inverse linear dependence of proteolysis rates with global thermodynamic stability (Figure 4.7). To determine the extent to which substrate stability affects rates of degradation, it is necessary to degrade more protein variants with different stabilities.

An advantage of using linear ankyrin repeat arrays as substrates is that it provides a means to relate the distance-dependence of local stability on rates of protein degradation. Lon protease processively degrades substrates from the initiation point of degradation, the degron tag. Because of this, we can ask how local changes in stability conferred from substitutions close to and far from the degron tag affect rates of degradation. We attempted to address this using linear four-repeat NRRC consensus ankyrin repeats; however we observe no detectable difference between destabilization at the protein N-terminus compared to the C-terminus.

It is possible that protein degradation by Lon protease is independent of local stability changes. However, the observation that degradation rates are dependent on total stability is at odds with a local stability-independent degradation mechanism. It is possible that four repeat NRRC consensus ankyrin is not long enough from N- to C-terminus for Lon protease to sense spatial stability differences in the folded state. Because we obtained intrinsic and interfacial contributions to the folding free energy, we can obtain the distribution of local folding free energies for consensus ankyrin repeats across of varying repeat number and identity (Figure 4.8). These distributions show that the addition of a fourth repeat from NRC to NRRC dramatically increases the stability of the internal R repeats. This results from the large favorable stability gained from the additional repeat interface. This increase in

stability is responsible for the observed decrease in protein degradation by Lon protease.

Next we examined the distribution of local folding free energies for consensus ankyrin repeats when destabilizing substitutions were made to either termini. The T4V substitution carries an intrinsic free energy of 5.63 ± 0.14 kcal*mol^{-1} (Chapter 2), which is slightly destabilizing compared to consensus ankyrin (Aksel et al., 2011). Substituting threonine with valine significantly decreased the stability of the interface between two substituted repeats, resulting in an interfacial free energy of -9.20 ± 0.20 kcal*mol^{-1} (compared to 12.54 kcal/mol for the R:R interface). When comparing the distribution of free energies for the substituted N*R*RC and NRR*C* ankyrin repeats (Figure 4.8), we observed that destabilizing the first two N-terminal repeats left the C-terminal repeat stability unchanged, compared to the parent NRRC. The same is true for the converse with substituted C-terminal repeats. However, the *in vitro* assay was not sensitive enough to differentiate between the two. Because of this, longer repeat arrays are required to measure different degradation rates resulting from local perturbations of thermodynamic stability.

It is well established that protein degradation by AAA^{+} proteases, of which Lon proteases is a member, requires ATP hydrolysis to power protein unfolding and translocation of the protein substrate through the

catalytic core. In fact, the amount of ATP hydrolyzed by AAA^+ proteases is directly related to global protein stability (Kenniston et al., 2003). This relationship implies that mechanical denaturation of the substrate proceeds through continual firing and re-firing of the ATPase motor. While stability and degradation rates are closely linked, we have yet to determine the extent to which this action occurs in a distant dependent manner.

4.5 Materials and Methods

4.5.1 Cloning, expression, and purification

To clone arrays of consensus ankyrin repeats, repeat variants, and the β20-degron tag (QLRSLNGEWRFAWFPAPEAV), we employed a complementary BamHI/BglII digested cloning site ligation strategy as described previously (Aksel et al., 2011) using 5'-phosphorylated synthetic DNA (Invitrogen) and cloned into a modified pET15b expression vector (Novagen). The gene for Lon protease was PCR amplified from of DH5α *E. coli* genomic DNA and cloned into a pBAD expression vector (ThermoFisher Scientific) with the gene encoding for an N-terminal His_{12} affinity tag separated by a TEV cleavage sequence (ENLYFQS).

E. coli BL21(DE3) were transformed with plasmids containing target protein genes, and were grown in Luria Broth at 37˚C to an OD_{600} of 0.6-0.8. Expression of repeat proteins was induced by addition of 1 mM IPTG, and the expression of Lon protease was induced by addition of 0.2% w/v arabinose. After further growth for 4-6 hours at 37˚C, cells were collected by centrifugation and frozen at -80˚C. Cell pellets containing repeat proteins were resuspended in 8 M urea, 1 M NaCl, and

25 mM Tris-HCl pH 8.0. Cell pellets containing Lon protease were resuspended in 1 M NaCl, and 25 mM Tris-HCl pH 8.0. Resuspended cells were lysed by sonication, and centrifuged to remove insoluble cell debris. The supernatant was loaded onto a nickel column (QIAGEN). After washing with the same resuspension buffer, bound protein was washed with 150 mM NaCl, 25 mM Tris-HCl pH 8.0. Repeat proteins were eluted with 0.5 M imidazole and His_{12}-tagged Lon protease was eluted with 1M imidazole. Pure protein-containing fractions were dialyzed overnight into the desired buffer to remove imidazole.

4.5.2 Lon protease degradation assay

Degradation assays were modified from previous work (Gur and Sauer, 2008) and carried out at 37˚C. Lon degradation buffer contained 50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 4 mM ATP, 20 mM phosphoenolpyruvate, and 50 U/mL Pyruvate Kinase (Roche). Degradation rate time points were collected by removal of reaction aliquots and quenching with 4x Coomassie stain buffer (40% glycerol, 240 mM Tris-HCl pH 6.8, 8% SDS, 0.04% bromophenol blue, 5% beta-mercaptoethanol). Gel samples were separated by Tris-Tricine-

SDS-PAGE (Schägger, 2006) and bands were analyzed by densitometric analysis using ImageQuantTL software (GE Healthcare).

4.5.3 Calculation of local folding free energies of ankyrin repeat proteins

Probability calculations were modified from previous work (Geiger-Schuller, 2016) using Ising parameters determined for consensus ankyrin repeats (Aksel et al., 2011) and ankyrin repeat substitutions (Chapter 2). Ising parameters have not been determined for N- and C-capping repeats containing T4V substitutions. Because of this, the difference between R repeats and R* repeats containing the T4V substitution was applied to the N- and C- intrinsic values, and the interfacial terms between substituted capping repeats is the same as R*-R*.

4.6 References

Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993 19, 349–360.

Bai, Y., Sosnick, T.R., Mayne, L., and Englander, S.W. (1995). Protein folding intermediates: native-state hydrogen exchange. Science 269, 192–197.

Baker, T.A., and Sauer, R.T. (2012). ClpXP, an ATP-powered unfolding and protein-degradation machine. Biochim. Biophys. Acta 1823, 15–28.

Gur, E., and Sauer, R.T. (2008). Recognition of misfolded proteins by Lon, a AAA+ protease. Genes Dev. 22, 2267–2277.

Gur, E., and Sauer, R.T. (2009). Degrons in protein substrates program the speed and operating efficiency of the AAA+ Lon proteolytic machine. Proc. Natl. Acad. Sci. 106, 18503–18508.

Gur, E., Vishkautzan, M., and Sauer, R.T. (2012). Protein unfolding and degradation by the AAA+ Lon protease. Protein Sci. Publ. Protein Soc. 21, 268–278.

Kenniston, J.A., Baker, T.A., Fernandez, J.M., and Sauer, R.T. (2003). Linkage between ATP consumption and mechanical unfolding during the protein processing reactions of an AAA+ degradation machine. Cell 114, 511–520.

Martin, A., Baker, T.A., and Sauer, R.T. (2007). Distinct static and dynamic interactions control ATPase-peptidase communication in a AAA+ protease. Mol. Cell 27, 41–52.

Park, C., and Marqusee, S. (2005). Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding. Nat. Methods 2, 207–212.

Schägger, H. (2006). Tricine-SDS-PAGE. Nat. Protoc. 1, 16–22.

Tripp, K.W., and Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. J. Mol. Biol. 365, 1187–1200.

Vieux, E.F., Wohlever, M.L., Chen, J.Z., Sauer, R.T., and Baker, T.A. (2013). Distinct quaternary structures of the AAA+ Lon protease control substrate degradation. Proc. Natl. Acad. Sci. 110, E2002–E2008.

Chapter 5

Polyglutamate- and polyaspartate-tagged protein affinity purification on hydroxyapatite

This chapter includes contributions from Timothy Barrick in the Barrick lab at The Johns Hopkins University.

5.1 Introduction

Scientists across all disciplines require pure reagents to successfully execute high quality experiments. In the molecular biosciences, experiments often require large amounts of high purity protein, often from recombinant sources. Numerous chromatographic methods have been developed to separate proteins of interest from other

undesirable species. For a review on expression and purification of proteins see (Structural Genomics Consortium et al., 2008).

The genetic fusion of affinity tags to target substrates allows for efficient separation from crude extract along a bound stationary phase. One of the most widely used affinity tag purification methods is immobilized metal-ion affinity chromatography (IMAC). In this method, a protein of interest is genetically fused to a hexameric histidine (His_6) affinity tag that binds a cationic metal, typically nickel or cobalt, allowing for separation from other soluble species. This technology, developed by Roche in the 1980's (Hochuli et al., 1988), uses high concentrations of the small molecule, imidazole, to compete with histidine-metal-ion binding interactions, allowing for the purification of recombinant protein.

After pure protein is obtained, the high concentrations of imidazole must be removed by dialysis. Imidazole presents a number of spectroscopic and stability complications. In large quantities, imidazole absorbs light at 280 nm, masking aromatic tryptophan and tyrosine absorbance used to determine protein concentration. Imidazole also absorbs at far-ultraviolet wavelengths, precluding circular dichroism spectroscopy to monitor secondary structure content. Recently, our lab has found that imidazole decreases protein stability, shifting the folding equilibrium towards the denatured state (data not shown). During

dialysis, proteins often aggregate, likely a result of the effect of imidazole on protein stability and solubility.

Here we present a novel technology that exploits the binding capacity of hydroxyapatite to a negatively charged amino acid affinity tag. Hydroxyapatite $Ca_5(PO_4)_3(OH)$, a major component of bone, binds negatively charged carboxyl residues via calcium-mediated interactions. A number of hydroxyapatite chromatographic protocols currently exist to purify antibodies, globular proteins, plasmid DNA, and acidic proteins (Bio-Rad), where proteins bound by negatively charged groups are eluted with increasing phosphate. However, a charged affinity tag exploiting these interactions has not been developed. In the present study, we design a genetically fused hexameric and dodecameric aspartic acid and glutamic acid sequences for their potential use in purifying recombinant proteins using hydroxyapatite.

5.2 Results

5.2.1 Ni-purified protein can be bound to hydroxyapatite and eluted with phosphate.

To determine the extent to which stretches of additional aspartic acid and glutamic acid residues affect affinity for hydroxyapatite, we have developed a series of affinity tags (Figure 5.1) that vary both in length and positioning. We use an engineered His_6-tagged enhanced green fluorescent protein (EGFP) fused to a consensus ankyrin three-repeat protein (NRC) as our target substrate. Consensus ankyrin repeat proteins are linear α-helical arrays that express well in *E. coli*, have relatively high thermodynamic stability (Aksel et al., 2011), and have been shown to increase the stability of adjacent proteins (Tripp and Barrick, 2007). EGFP-NRC has been previously used in the lab, overexpresses in large amounts in *E. coli*, and is stable and highly soluble at room temperature (data not shown). EGFP contains a chromophore that gives the protein a green color under ambient light conditions. This makes EGFP ideal for affinity-purification on an

immobile white hydroxyapatite background, allowing for simple detection by eye.

For standard MAC purification, the addition of 6 histidines is sufficient to immobilize soluble proteins on stationary nickel agarose beads. Here we employ the same strategy, genetically fusing 6 aspartic or glutamic acid residues to the C-terminus of our target protein (Figure 5.1A). To determine the extent to which additional residues affect hydroxyapatite binding, we have also built 12-residue affinity tags (Figure 5.1A).

To test hydroxyapatite binding, we first expressed and purified our genetically engineered protein by Ni-affinity purification using the N-terminal His_6 tag. After this initial purification, the target 39.2 kDa protein is roughly 50% pure (Figure 5.2). After removal of imidazole by dialysis, the Ni-purified protein was loaded onto hydroxyapatite resin at neutral pH, and was washed with increasing concentration NaCl and Na_3PO_4. During the loading and washing procedure, it was clear that the EGFP was retained on the top of the column. The final bound EGFP fusion protein was eluted from the column with high concentration Na_3PO_4.

SDS-PAGE analysis reveals the extent to which the affinity tags bind hydroxyapatite (Figure 5.2, Table 5.1), as well as the degree of purification. Increasing the tag length from 6 to 12 residues allows the

protein remain bound to the hydroxyapatite resin through more stringent washing conditions. This allows for the removal of soluble contaminating proteins, increasing the purity of the eluted protein. Furthermore, a 12-residue glutamic acid affinity tag has a higher affinity for hydroxyapatite than an equal length aspartic acid tag, though the 6-residue asp and glu tags appeared to have similar affinities to one another. We have outlined the upper limit of wash buffer composition for each of the four tags (Table 5.1).

5.2.2 Hydroxyapatite purification directly from crude lysate

Our previous work has utilized hydroxyapatite affinity purification as a second-pass procedure, starting with relatively pure proteins. To test whether hydroxyapatite affinity purification can be performed directly on crude lysate (Figure 5.3A), we passed soluble *E. coli* cell lysate containing overexpressed H_6-EGFP-NRC-E_{12} protein over immobile hydroxyapatite. We observe similar affinity from crude lysate as we did with nickel-purified protein. This protein can then be further purified over Ni-NTA resin via the N-terminal His_6 tag (5.3B).

5.2.3 Hydroxyapatite purification in high concentrations of imidazole

This sample of >95% purity (by SDS-PAGE) contains high concentrations of imidazole. Rather than removing imidazole by dialysis, we find that tagged-protein samples in high imidazole concentration can bind directly hydroxyapatite without first removing imidazole (5.3C), and then eluting with a high phosphate step. This imidazole removal step appears to further increase purity. Although the phosphate must be removed by dialysis or dilution, phosphate is not known to destabilize proteins. Thus, dialysis to remove phosphate is expected to be less problematic than dialysis to remove imidazole.

5.2.4 Immunodetection of polyglutamate affinity tag

An additional potential application of the E6 and E12 tags is for use in immunological detection and immunoprecipitation. Because polyglutamate is a post-translational modification of microtubules (Boucher et al., 1994), antibodies against Glu6 are commercially available. Here we tested whether we could 6-residue and 12-residue polyglutamate tags by Western blotting (Figure 5.3D). Target proteins

containing only a His_6 tag, both His_6 and Glu_6 tags, and His_6 and Glu_{12} tags were probed by α-His and α-Glu antibodies. All proteins that contained His_6 tags were detected by Western blot using α-His antibodies, whereas only proteins that contained polyglutamic acids were detected by α-Glu antibodies.

5.2.5 Purification of adenylate kinase using Asp and Glu tags.

To test whether Asp- and Glu-tag purification could be extended to other proteins, we fused 12-residue aspartic acid and 12-residue glutamic acid tags to the N-terminus and C-terminus of adenylate kinase (AdK). As with the EGFP-NRC construct, we observed tight binding and retention of tagged AdK to hydroxyapatite (Figure 5.4B), whereas the His_6-only tagged adenylate kinase eluted through in the flow-through, and in mild wash conditions (Figure 5.4A). For N-terminally tagged adenylate kinase, we observe a low molecular weight species that co-purifies with tagged AdK the first round of Ni-affinity purification. This is possibly protein whose N-terminal polyglutamate and polyaspartate tag was either not translated or removed by proteolysis. These lower molecular weight species wash off the hydroxyapatite before intact tagged protein is eluted.

5.2.6 Polyglutamate and polyaspartate affinity tagged AdK retains enzymatic activity.

The use of adenylate kinase allows us to determine if polyglutamic acid and polyaspartic acid affinity tags impairs catalytic efficiency. To determine adenylate kinase activity, we employ the forward reaction mechanism of adenylate kinase, converting ATP and AMP into 2 ADP (Hamada et al., 1985). In this coupled reaction scheme, pyruvate kinase converts ADP and phosphoenolpyruvate to pyruvate and ATP. Pyruvate and NADH are then converted to NAD^+ and lactate by lactate dehydrogenase. This reaction is followed spectroscopically at 340 nm, the wavelength that NADH, but not NAD, absorbs light.

For each concentration of ATP substrate, we monitored absorbance at 340 nm over time (Figure 5.5A) to determine initial rates of reaction (V_0). These initial rates are then plotted as a function of ATP concentration and fit with a Michaelis-Menten model to determine the kinetic parameters V_{max} and K_M (Figure 5.5B, Table 5.2). We observe similar kinetic parameters (k_{cat} and V_{max}) for the N- and C-terminally His_6-tagged adenylate kinases. The H_6D_{12} and H_6E_{12} constructs have

similar K_m values to the H_6 constructs, but they have an approximate two-fold decrease in V_{max} values (Figure 5.6, Table 5.2). This indicates that the presence of multiple aspartic acid and glutamic acid residues at either N- or C-terminus decreases the catalytic efficiency of *E. coli* adenylate kinase. Therefore, we recommend placing an internal protease site (e.g., TEV, precision, or fibronectin) between the affinity tag and purified proteins whose binding and function will be characterized biochemically.

5.3 Figures and Tables

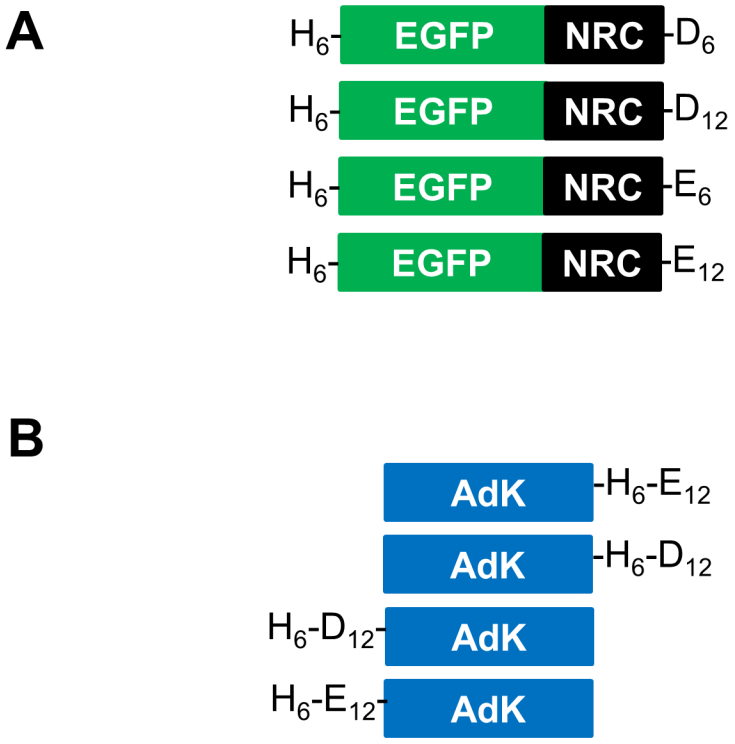


Figure 5.1 Schematic of polyaspartate and polyglutamate tagged substrates. (A) Hexameric and dodecameric polyaspartate and polyglutamate affinity tags are fused to the C-terminus of His_6-EGFP-NRC to test binding to hydroxyapatite. (B) Dodecameric polyaspartate and polyglutamate are fused to both the N- and C-termini of *E. coli* adenylate kinase, adjacent to the His6 tag, to test terminal effect on binding to hydroxyapatite.

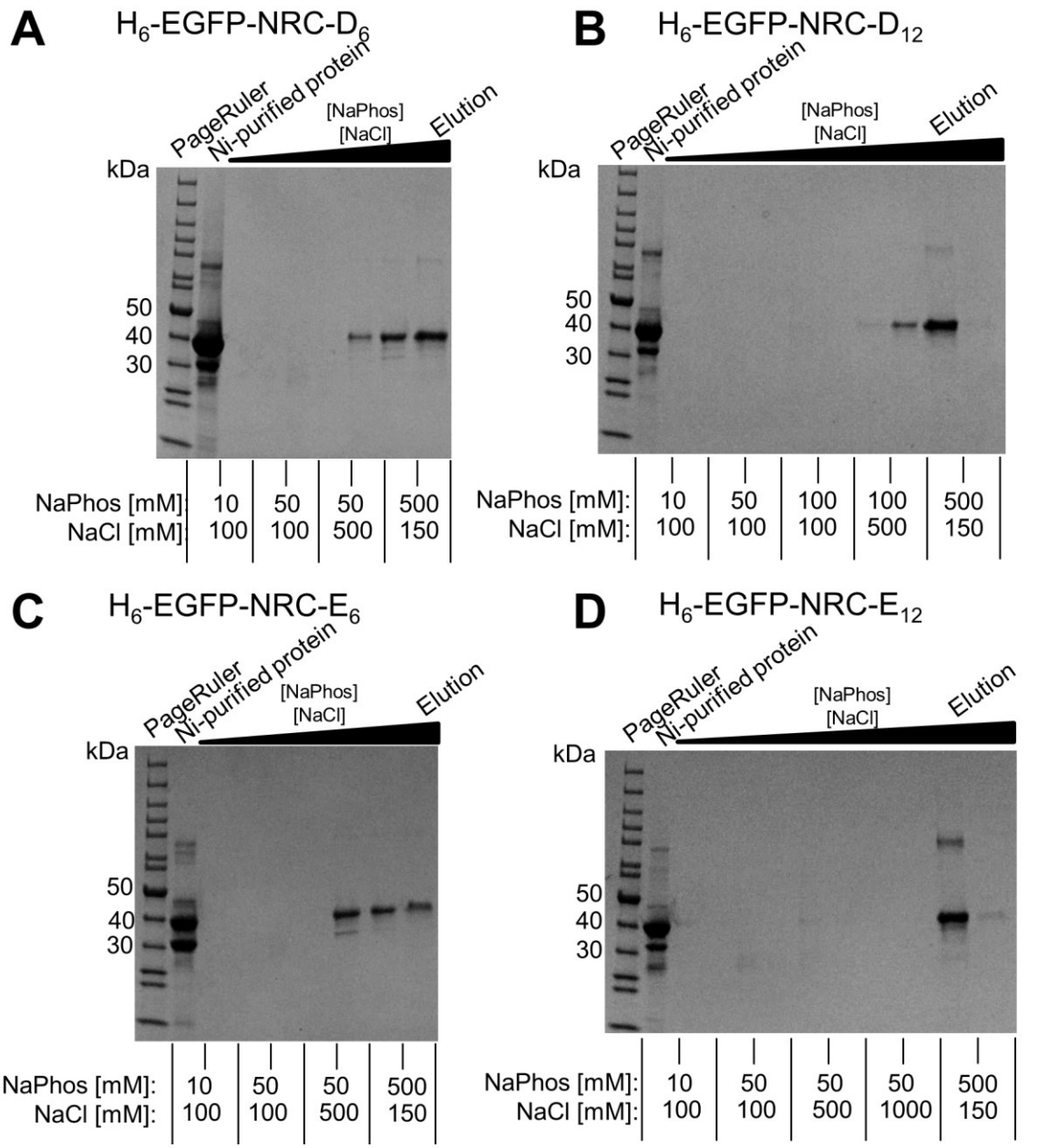


Figure 5.2 Polyglutamate and polyaspartate affinity tag binds hydroxyapatite. SDS-PAGE images of purification of EGFP-NRC with C-terminal polyaspartate (A) hexameric and (B) dodecameric affinity tags, and polyglutamate (C) hexameric and (D) dodecameric affinity tags. Protein sample was previously purified by Ni-His affinity chromatography (lane 2) and loaded on Hydroxyapatite resin. Sample was washed with increasing amounts of NaCl and Na_3PO_4 before elution with 500 mM Na_3PO_4.

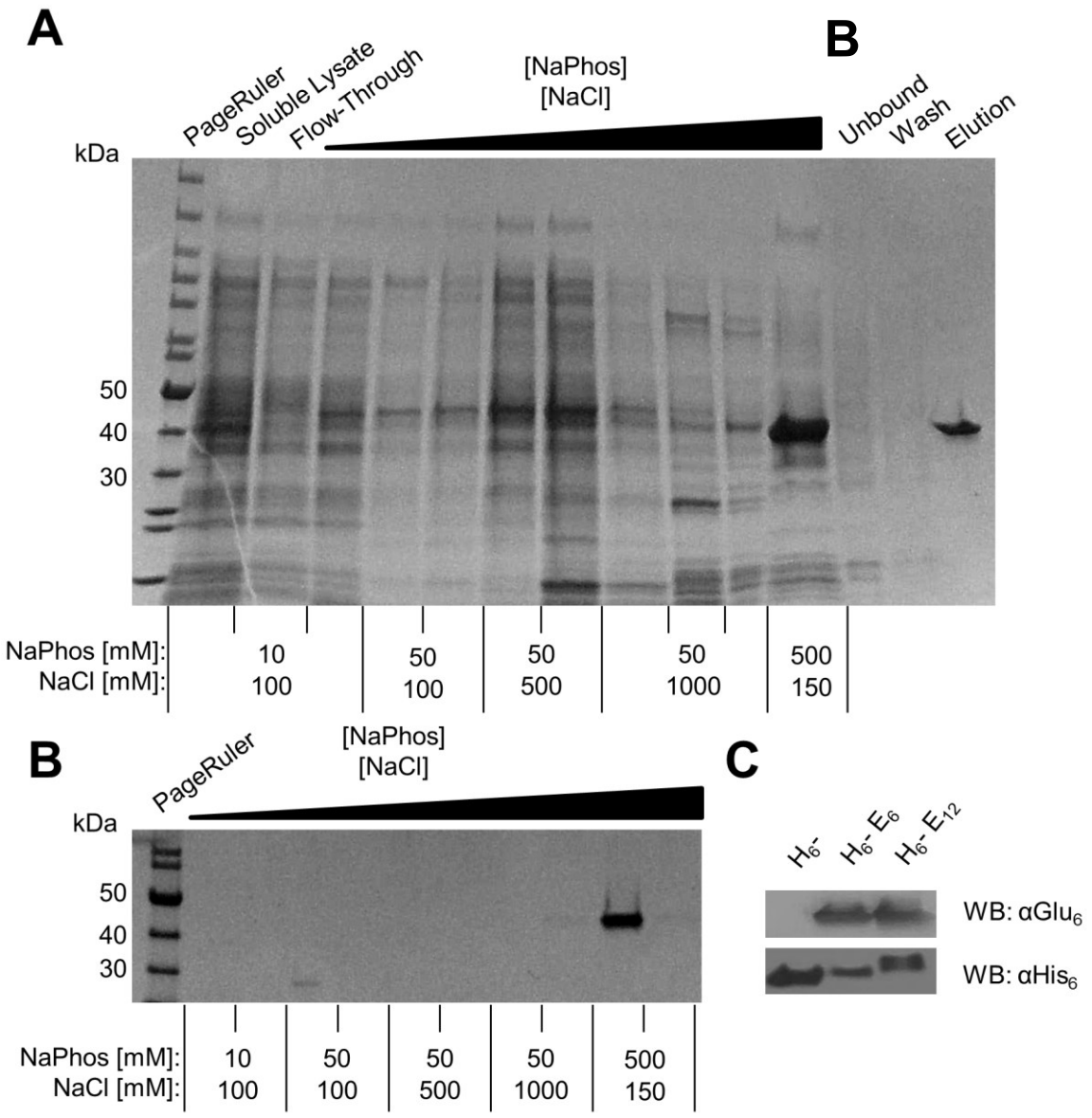


Figure 5.3 Dual-purification of dodecameric polyglutamate affinity-tagged protein and detection by Western blot. (A) SDS-PAGE image of hydroxyapatite purification from cell lysate. Soluble lysate from E.coli overexpression of H_6-EGFP-NRC-E_{12} was loaded onto hydroxyapatite resin and washed with increasing amounts of Na_3PO_4 before elution with 500 mM Na_3PO_4. (B) Hydroxyapatite-eluted protein was loaded onto Ni-NTA resin, washed with buffer, eluted with 500mM imidazole, and analyzed by SDS-PAGE. (C) Pure protein in 500mM imidazole from (B) binds hydroxyapatite resin, allowing for quick removal of imidazole. Bound protein was washed with increasing amounts of Na_3PO_4 before elution with 500 mM Na_3PO_4. (D) Polyglutamate tag is a useful epitope tag for Western blot detection. Western blot of H_6-EGFP-NRC (left) and polyglutamate-tagged H_6-EGFP-NRC (middle and right) probed with α-Glu_6 (top) and α-His_6 antibody (bottom).

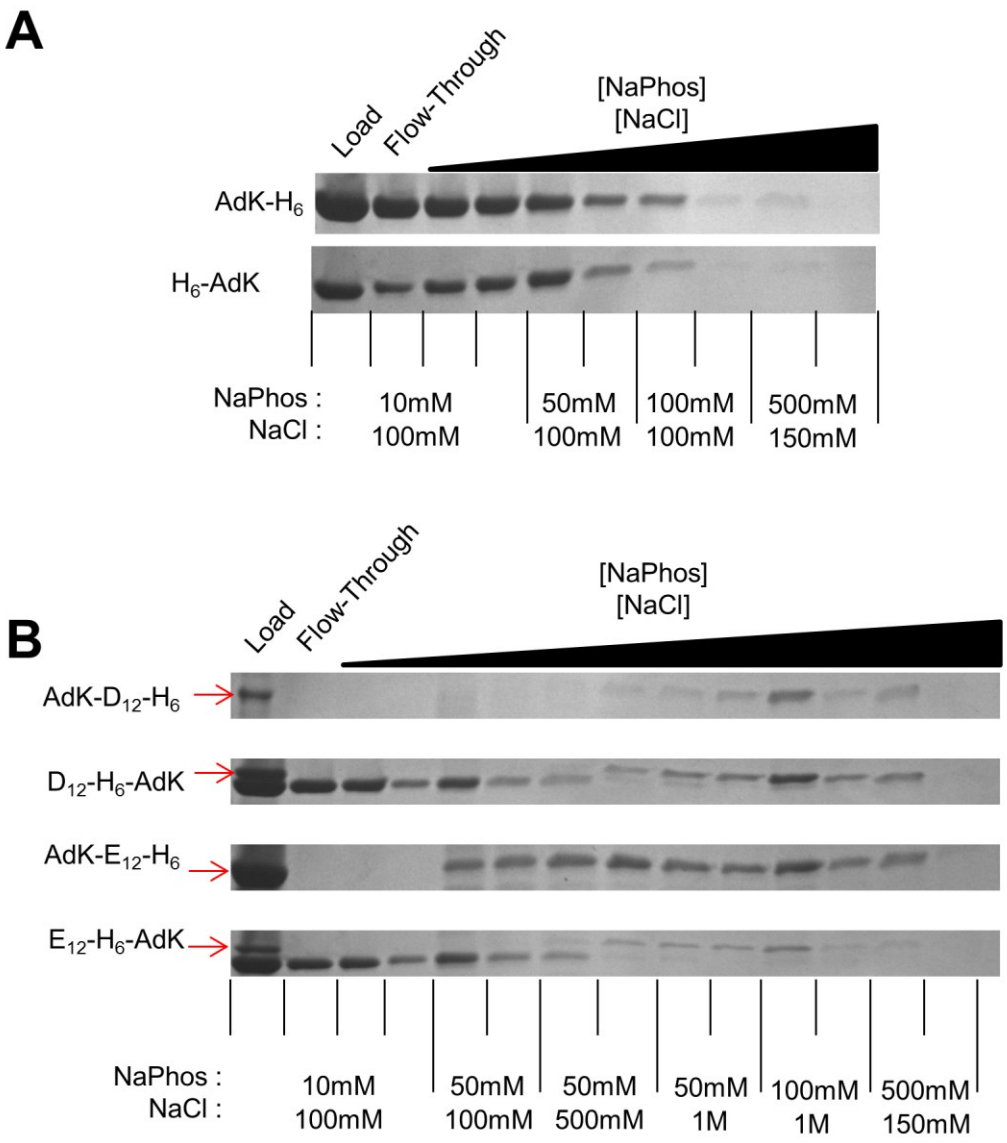


Figure 5.4 Polyaspartate and polyglutamate affinity for hydroxyapatite is independent of fusion to protein N- or C-termini. Ni-purified protein was loaded onto column, washed with increasing NaCl and Na3PO4 before elution with 500 mM Na3PO4. (A) SDS-PAGE analysis of purification of *E. coli* adenylate kinase with His_6-tag fused to C- (top) or N- (bottom) terminus on hydroxyapatite. (B) SDS-PAGE analysis of hydroxyapatite purification of *E. coli* adenylate kinase with D_{12}- and E_{12}-tags fused to C- or N- termini. Lower molecular weight untagged AdK for the N-terminally tagged construct washes off the hydroxyapatite before intact tagged protein.

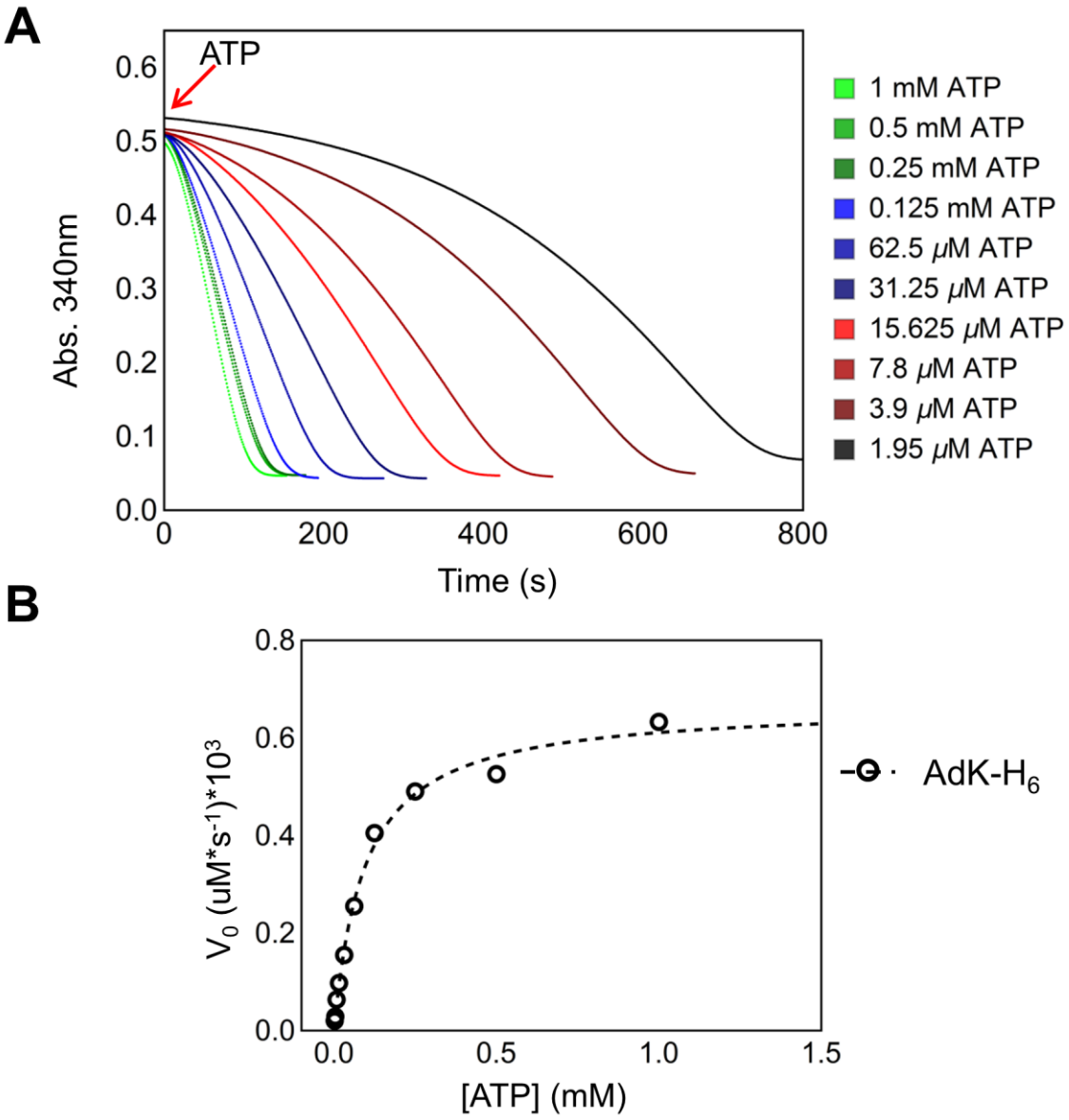


Figure 5.5 Analysis of AdK-H_6 reaction kinetics. (A) NADH absorbance was monitored at 340 nm at different starting concentrations of ATP. (B) Initial rates (open circles) from (A) were determined for each concentration of ATP, and fit with a Michaelis-Menten model ($V_0 = V_{max}[S]/(K_M + [S])$ shown in dashed lines. Initial rates increase with concentration of ATP until substrate saturation is reached.

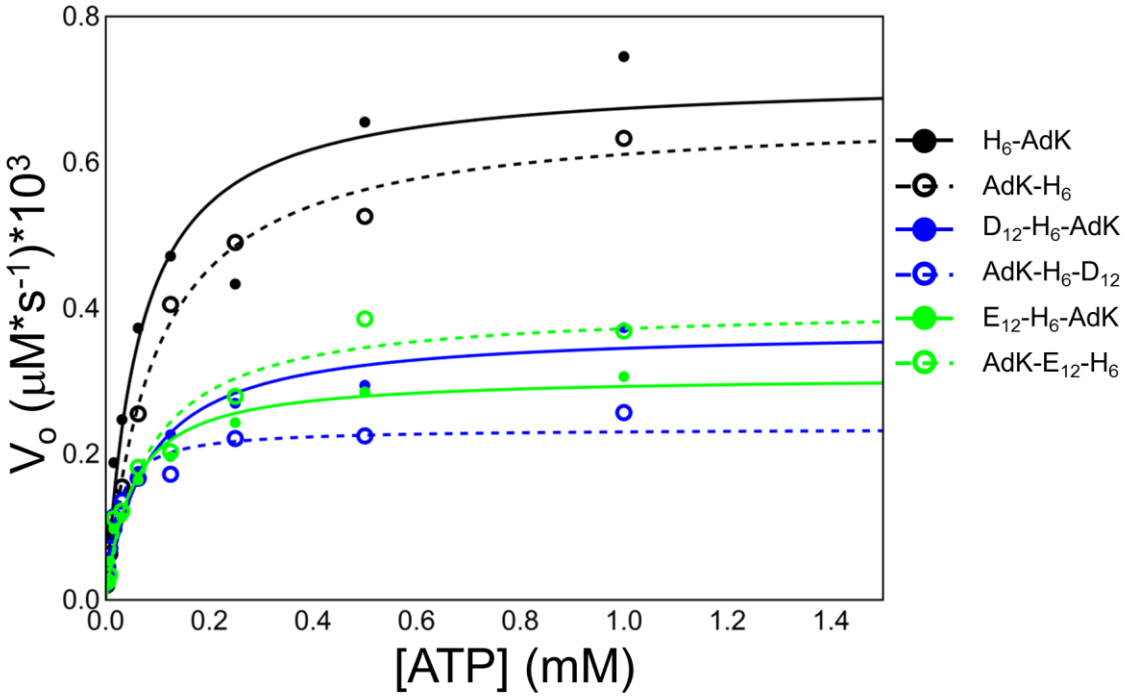


Figure 5.6 Terminal fusion of polyaspartate and polyglutamate affinity tags to *E. coli* adenylate kinase retain activity, albeit with a reduced Vmax value. Initial rates determined for N-terminal (solid circles) and C-terminal (open circles) H_6 (black), H_6D_{12} (blue), and H_6E_{12} (green) tagged *E. coli* adenylate kinase are fit with a Michaelis-Menten model ($V_0 = V_{max}[S]/(K_M + [S])$ shown in solid and dashed lines. H_6-only tagged adenylate kinase shows a higher V_{max} than D_{12}- and E_{12}-tagged adenylate kinase.

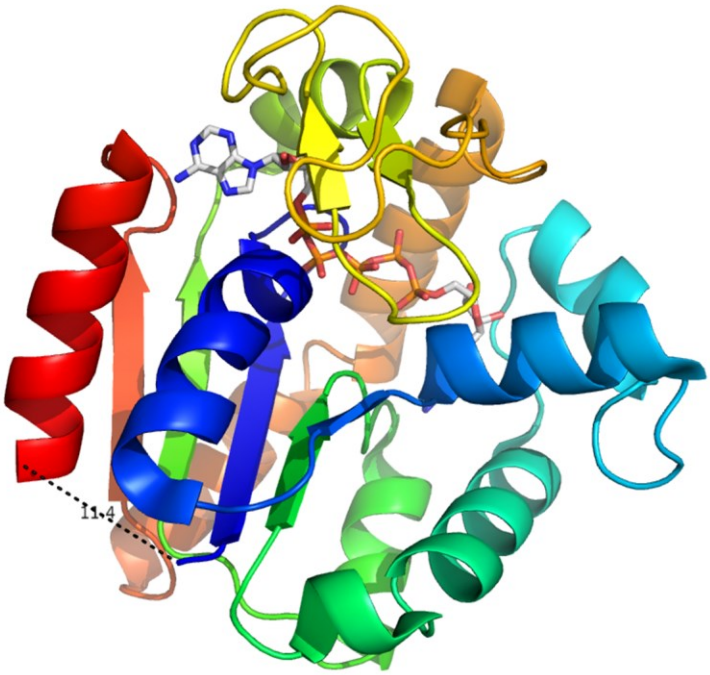


Figure 5.7 *E. coli* adenylate kinase termini are near each other in the folded structure. (PDB: 4X8O) Ribbon model of *E. coli* adenylate kinase in the closed "substrate" bound conformation (Ap5A inhibitor shown in sticks bound in the active site). N-terminus (blue) is 11.4 Å from the C-terminus (red).

Table 5.1 Wash tolerances for hydroxyapatite affinity column chromatography of Asp- and Glu-tagged EGFP-NRC constructs.

Tag:	Buffer Component:	
	NaCl	Na_3PO_4
D_6	500 mM	50 mM
D_{12}	500 mM	100 mM
E_6	500 mM	50 mM
E_{12}	1 M	50 mM

In each case, the tag was fused to the C-terminus of EGFP-NRC (see Figure 5.1).

Table 5.2 Kinetic parameters for N-terminally and C-terminally tagged *E. coli* adenylate kinase.

N-term. Tagged AdK	K_M (μM)	V_{max} $*10^3$ (μM/s)	k_{cat} $V_{max}/[E]$ (s^{-1})
H_6	0.064 ± 0.016	0.717 ± 0.127	179 ± 12.7
H_6D_{12}	0.078 ± 0.017	0.372 ± 0.051	92.9 ± 5.1
H_6E_{12}	0.050 ± 0.008	0.307 ± 0.033	76.8 ± 3.3

C-term. Tagged AdK	K_M (μM)	V_{max} (μM/s) $*10^3$	k_{cat} $V_{max}/[E]$ (s^{-1})
H_6	0.094 ± 0.009	0.669 ± 0.049	167.1 ± 4.9
H_6D_{12}	0.020 ± 0.004	0.235 ± 0.026	58.8 ± 2.6
H_6E_{12}	0.081 ± 0.020	0.402 ± 0.071	100.5 ± 7.1

Kinetic parameters obtained from Michaelis-Menten equation fit to initial rates determined from spectroscopic absorbance at 340nm. Standard error obtained from the nonlinear least squares variance-covariance matrix assuming parameter uncertainties to be distributed normally.

5.4 Discussion

Here we present a novel affinity chromatography method for the purification of recombinant proteins. This method relies on the binding of hydroxyapatite to negatively charged aspartic acid and glutamic acid residues. We have genetically engineered 6-residue and 12-residue affinity tags to an EGFP protein as well as *E. coli* adenylate kinase, and find tagged proteins to bind to hydroxyapatite, to be retained through high salt washes, and to be eluted by high concentrations of sodium phosphate. The affinity tag can be located on either N- or C-terminus, and provides a rapid means to purify proteins.

Moreover, the method described here is orthogonal to standard His_6 affinity purification, and can be used in conjunction with nickel chromatography. Importantly, tagged recombinant protein binds hydroxyapatite in the presence of high concentration imidazole, allowing imidazole to be removed from protein samples without a lengthy dialysis step. This avoids complications due to imidazole-induced protein destabilization, such as aggregation, and reduced yield.

We present a single instance where the presence of polyglutamate and polyaspartate at either N- or C-terminus of an enzyme, *E. coli* adenylate kinase, reduces the catalytic activity of the enzyme. When

examining the crystal structure of adenylate kinase in the bound closed conformation, we observe a close spatial proximity of the N-terminus relative to the C-terminus (Figure 5.7). Adenylate kinase binds negatively charged nucleotides in order to perform its catalytic function. It is possible that the introduction of a 12-residue stretch of negatively charged glutamic acid or aspartic acid residues at either terminus reduces the catalytic activity of adenylate kinase through charge-charge repulsion, although this would seem more likely to affect k_{cat} than V_{max}. It is also possible that the tag impairs conformational changes believed to be important for the catalytic cycle (Kerns et al., 2015; Schrank et al., 2013).

Therefore for instances where the target fusion protein is involved in catalysis or other protein-mediated interactions, we recommend removal of the fusion tag. This can be easily done through cleavage by a site-specific protease with an engineered intervening cleavage site, like TEV protease with a TEV cleavage site.

Protein purification resins are relatively expensive. Currently Ni-NTA resins cost over $11 (USD) per 1 mL of a 50% resin slurry (Qiagen). In comparison, hydroxyapatite is significantly less expensive than Ni-NTA agarose, at roughly $2 (USD) per 0.5g dry powder (yielding 1 ml of resin slurry). Furthermore, hydroxyapatite can be readily reused. The cost

savings for this resin increases when compared to ion-exchange resins used on FPLC instrumentation.

The ability to obtain high purity recombinant protein is vital for any lab in biological sciences. It is our hope that the development of the hydroxyapatite purification technology using polyglutamate and polyaspartate affinity tags will increase capabilities of lab sciences world-wide.

5.5 Materials and Methods

5.5.1 Cloning, expression, and purification

The gene for EGFP was a gift from the lab of Madeline Shea (University of Iowa). Consensus ankyrin repeat proteins were cloned as previously described (Aksel et al., 2011). The gene for *E. coli* adenylate kinase was PCR amplified from of DH5α *E. coli* genomic DNA. All genes were subcloned into a modified pET15b expression vector (Novagen). 5'-phosphorylated synthetic DNA coding for hexameric and dodecameric polyglutamate and polyaspartate residues (Invitrogen) were ligated into appropriate restriction sites.

E. coli BL21(DE3) were transformed with plasmids containing repeat protein genes, and were grown in Luria Broth at 37˚C to an OD_{600} of 0.6-0.8. Expression was induced by addition of 1 mM IPTG. After further growth for 4-6 hours at 37˚C, cells were collected by centrifugation and frozen at -80˚C. Cell pellets were resuspended in 150 mM NaCl, and 25 mM Tris-HCl pH 8.0, lysed by sonication, and centrifuged to remove insoluble cell debris. The supernatant was loaded onto a nickel-NTA resin column (QIAGEN). After washing with five column volumes of resuspension buffer, bound protein was eluted with 0.5 M imidazole, 150 mM NaCl, 25 mM Tris-HCl pH 8.0. Pure protein-

containing fractions were dialyzed overnight into the desired buffer to remove imidazole.

5.5.2 Hydroxyapatite column preparation and purification

Hydroxyapatite (Sigma 289396) was washed in triplicate with 10 mM Na_3PO_4 pH 7.0 to remove contaminants. Protein-containing samples were loaded onto each column bed and washed with 10 column volumes of buffer containing 10mM Na_3PO_4 pH 7.0. Bound protein was washed as described with increasing concentration of NaCl and Na_3PO_4 all at pH 7.0, and eluted with 500 mM Na_3PO_4 pH 7.0. To remove any excess bound protein, hydroxyapatite was washed with 10 column volumes of 500 mM Na_3PO_4 pH 7.0 and re-equilibrated with 10 mM Na_3PO_4 pH 7.0. Protein samples were separated by SDS-PAGE over 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad 4561096).

5.5.3 Western Blot Detection

Samples containing pure protein were separated by SDS-PAGE over 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels (Bio-Rad 4561096). Proteins were then transferred to nitrocellulose membrane using a Trans-Blot® Turbo™ Transfer System (Bio-Rad). Membranes

containing protein samples were blocked with 5% bovine serum albumin (BSA) in 1x TBST (tris-buffered saline tween). Membranes were washed five times with 1x TBST containing 0.5% BSA, and probed with either α-Glu (AdipoGen 25B-0030-C050) or α-His (Abcam 125262) diluted 1:2000 in 1x TBST containing 0.5% BSA. Membranes were washed five times with 1x TBST containing 0.5% BSA, and positive bands were detected using HRP-conjugated anti-Rabbit in 1x TBST containing 0.5% BSA.

5.5.4 Enzyme assay and kinetic study

The assay for *E. coli* adenylate kinase enzyme activity was adapted from previous work (Hamada et al., 1985). Reactions were carried out in 50 mM HEPES, 0.5 mg/mL BSA, 20 mM $MgCl_2$, 100 mM KCl, 1 mM phosphoenolpyruvate, 1 mM ATP, 0.1 mM NADH, 10 U/mL pyruvate kinase (Roche), 10 U/mL lactate dehydrogenase (Sigma). Reactions were monitored at room temperature by following absorbance at 340 nm (AVIV Associates: Lakewood, NJ).

5.6 References

Aksel, T., Majumdar, A., and Barrick, D. (2011). The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993 19, 349–360.

Boucher, D., Larcher, J.C., Gros, F., and Denoulet, P. (1994). Polyglutamylation of tubulin as a progressive regulator of *in vitro* interactions between the microtubule-associated protein Tau and tubulin. Biochemistry (Mosc.) 33, 12471–12477.

Hamada, M., Sumida, M., Kurokawa, Y., Sunayashiki-Kusuzaki, K., Okuda, H., Watanabe, T., and Kuby, S.A. (1985). Studies on the adenylate kinase isozymes from the serum and erythrocyte of normal and Duchenne dystrophic patients. Isolation, physicochemical properties, and several comparisons with the Duchenne dystrophic aberrant enzyme. J. Biol. Chem. 260, 11595–11602.

Hochuli, E., Bannwarth, W., Döbeli, H., Gentz, R., and Stüber, D. (1988). Genetic Approach to Facilitate Purification of Recombinant Proteins with a Novel Metal Chelate Adsorbent. Nat. Biotechnol. 6, 1321–1325.

Kerns, S.J., Agafonov, R.V., Cho, Y.-J., Pontiggia, F., Otten, R., Pachov, D.V., Kutter, S., Phung, L.A., Murphy, P.N., Thai, V., et al. (2015). The energy landscape of adenylate kinase during catalysis. Nat. Struct. Mol. Biol. 22, 124–131.

Schrank, T.P., Wrabl, J.O., and Hilser, V.J. (2013). Conformational heterogeneity within the LID domain mediates substrate binding to *Escherichia coli* adenylate kinase: function follows fluctuations. Top. Curr. Chem. 337, 95–121.

Structural Genomics Consortium, China Structural Genomics Consortium, Northeast Structural Genomics Consortium, Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., Knapp, S., et al. (2008). Protein production and purification. Nat. Methods 5, 135–146.

Tripp, K.W., and Barrick, D. (2007). Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. J. Mol. Biol. 365, 1187–1200.

Vita

Kevin Andrew Sforza was born on October 31, 1987 in Trenton, New Jersey. He attended Alexander Elementary School in Hamilton Square before moving to Toms River, New Jersey. There he attended North Dover Elementary and Joseph A. Citta Elementary schools, and Toms River Intermediate West. He graduated from Toms River High School North in 2006. Kevin attended the University of Delaware in Newark, Delaware and performed research in the laboratory of Dr. Zhihao Zhuang. He graduated in 2010 with a bachelor of science with honors in biochemistry and a minor in anthropology. From 2010 to 2012, Kevin worked as a Scientist I at the custom antibody company, SDIX, Inc. in Newark, DE. In 2012, he joined the program in Cell, Molecular, Developmental Biology, and Biophysics at The Johns Hopkins University in Baltimore, Maryland to pursue his Ph.D. in Biological Sciences. He completed his dissertation research in the laboratory of Dr. Douglas E. Barrick in the summer of 2017.